DOCUMENT RESUME

ED 114 406 95 TM 004 856

AUTHOR Horst, Donald P.; And Others
TITLE Measuring Achievement Gains in Educational Projects.
INSTITUTION RMC Research Corp., Los Altos, Calif.
SPONS AGENCY Office of Education (DHEW), Washington, D.C. Office of Planning, Budgeting, and Evaluation.
REPORT NO RMC-UR-243
PUB DATE Oct 74
NOTE 122p.; This document is superseded by ED 106 376, A Practical Guide to Measuring Project Impact on Student Achievement. Monograph Series on Evaluation in Education No. 1

EDRS PRICE MF-$0.76 HC-$5.70 Plus Postage
DESCRIPTORS *Achievement Gains; Analysis of Covariance; *Cognitive Measurement; Data Analysis; Data Collection; Decision Making; Educational Researchers; Evaluation Methods; *Guides; Information Dissemination; Measurement Techniques; *Models; Program Development; *Program Evaluation; Research Methodology; Research Problems; Selection; Standardized Tests; Testing

ABSTRACT

Directors of educational projects need to be aware of the consequences their decisions may have for evaluation and appreciate the need for working closely with their evaluators from the earliest planning stage. Attempting to address the needs of project directors and evaluators, this guidebook deals with one central aspect of project evaluation--measuring congnitive achievement gains. Its purpose is to provide the tools needed to conduct technically sound, interpretable evaluation studies. It covers the entire evaluation process from the administrative decisions in selecting an evaluation design to the details of collecting, analyzing, and reporting the data. After the introduction, Chapter 2 describes 12 hazards commonly encountered in evaluations which may invalidate otherwise sound studies. The hazards are discussed, and ways to avoid the hazard are outlined. Chapter 3 presents a procedural guide, in decision-tree form, for selecting a suitable evaluation model given a particular set of constraints. Chapter 4 presents the five evaluation models referred to in Chapter 3. Each model is summarized describing its characteristics, strengths, weaknesses, and considerations relating to its implementation. Chapter 5 deals with the details of data collection and Chapter 6 with summarizing and reporting of impact data. Appendices contain characteristics of some commonly used standardized tests and analysis of covariance worksheets. (RC)

Documents acquired by ERIC include many informal unpublished materials not available from other sources. ERIC makes every effort to obtain the best copy available. Nevertheless, items of marginal reproducibility are often encountered and this affects the quality of the microfiche and hardcopy reproductions ERIC makes available via the ERIC Document Reproduction Service (EDRS). EDRS is not responsible for the quality of the original document. Reproductions supplied by EDRS are the best that can be made from the original.

ED114406

RMC Report
UR-243

MEASURING ACHIEVEMENT GAINS IN EDUCATIONAL PROJECTS

Donald P. Horst
G. Kasten Tallmadge
Christine T. Wood

U.S. DEPARTMENT OF HEALTH, EDUCATION & WELFARE
NATIONAL INSTITUTE OF EDUCATION

THIS DOCUMENT HAS BEEN REPRODUCED EXACTLY AS RECEIVED FROM THE PERSON OR ORGANIZATION ORIGINATING IT. POINTS OF VIEW OR OPINIONS STATED DO NOT NECESSARILY REPRESENT OFFICIAL NATIONAL INSTITUTE OF EDUCATION POSITION OR POLICY

October 1974

Prepared for
U. S. DEPARTMENT OF HEALTH, EDUCATION, AND WELFARE
Office of Education/Office of Planning, Budgeting, & Evaluation

The research reported herein was performed pursuant to a contract with the Office of Education, U. S. Department of Health, Education, and Welfare. Contractors undertaking such projects under Government sponsorship are encouraged to express freely their professional judgment in the conduct of the project. Points of view or opinions stated do not, therefore, necessarily represent official Office of Education position or policy.

RMC Research Corporation
Los Altos, California

TM 004 856

ACKNOWLEDGEMENTS

The authors of this report are indebted to many for large amounts of help. Most deserving of special thanks are the members of a special Advisory Panel who reviewed and commented on an early draft. Members of the panel included James D. Bailey of the Los Angeles Unified School District, Donald T. Campbell of Northwestern University, Mark M. Greene of the Northwest Regional Educational Laboratory, Roger J. Lulow of the Ohio State Department of Education, and John Stindt of the Highland Park, Michigan School District. Their contributions are reflected throughout the guidebook and have added greatly to its comprehensiveness, accuracy, and readability.

Edward B. Glassman of the U.S. Office of Education's Office of Planning, Budgeting, and Evaluation was the Project Officer responsible for preparation of the guidebook. He deserves special thanks for his own thoughtful review of the document and for collecting additional ideas from his professional colleagues. Paul Horst helped greatly with those portions of the guidebook dealing with regression analysis and related issues. Paul Wortman of Northwestern also reviewed draft materials and provided useful suggestions. Finally, the authors are grateful to Diane Jones and Lora Caldwell for their clerical assistance in preparing this manuscript and for their patience and good humor in dealing with seemingly endless revisions.

D.P.H.
G.K.T.
C.T.W.

TABLE OF CONTENTS

	Page
ACKNOWLEDGEMENTS	iii
TABLE OF CONTENTS	v
LIST OF FIGURES	vii
I. INTRODUCTION	1
II. COMMON HAZARDS IN EVALUATION	8
III. A PROCEDURAL GUIDE FOR MODEL SELECTION	33
IV. EVALUATION MODELS	48
V. GETTING THE DATA (TESTING AND RECORDING)	76
VI. ANALYZING THE DATA AND REPORTING THE RESULTS	85
APPENDIX A	
Characteristics of Commonly Used Standardized Tests	92
APPENDIX B	
Analysis of Covariance Worksheets	108
REFERENCES	115

LIST OF FIGURES

Figure		Page
1	Decision tree for selecting evaluation models . . .	47
2	Score distributions with treatment effect independent of pretest status	61
3	Score distributions with treatment effect inversely related to pretest status	62
4	Sample data form	81

I. INTRODUCTION

Purpose and Scope

The evaluation of any special instructional project is affected by decisions made at all levels of project administration and at all stages of planning and implementation. All too often an evaluation specialist is brought in after a project is well under way only to find that actions have already been taken which make it difficult, if not impossible, to perform any kind of meaningful impact assessment.

To avoid this clearly undesirable situation, directors of educational projects need to be aware of the consequences their decisions may have for evaluation and to appreciate the need for working closely with their evaluators from the earliest planning stage. This guidebook attempts to address the needs of project directors as well as evaluators, and the next section of this chapter specifically designates certain sections as "recommended reading" for project directors.

The guidebook deals with only one central aspect of project evaluation, measuring cognitive achievement gains. It is not concerned with project costs or with any affective benefits which project participants may accrue. Neither does it address any such "process" variables as how well the objectives of the project were stated, how well the needs of the children were assessed, or how closely teachers followed prescribed instructional strategies. The entire focus is on obtaining as clear and unambiguous an answer as possible to the question, "How much more did pupils learn by participating in the project than they would have learned without it?"

The guidebook is the result of a search by the authors for effective compensatory reading and mathematics projects (Tallmadge, 1974). The search encompassed some 2,000 projects, all of which had received some form of "official" recognition for success. Of the 2,000, only six could be found which, under close scrutiny, were able to meet the selection criteria of effectiveness, cost, availability, and replicability

established for this search (see Foat, 1974). Most discouraging, however, was the fact that not one of the evaluations provided acceptable evidence regarding project success or failure. In all cases, problems in conducting and reporting the evaluations rendered the results inconclusive. Obviously, practical considerations prevent school evaluators from doing controlled, laboratory experiments, but many of the problems in current evaluation practices could be avoided with little or no increase in cost or effort. The rigor of laboratory experimentation may be beyond reach, but the state-of-the-art can be greatly improved without placing unrealistic demands on schools or evaluation resources.

The purpose of this guidebook is to provide those concerned with project evaluation with the basic tools they need to conduct technically sound, interpretable evaluation studies. Every effort has been made to minimize the amount of technical sophistication required of users of the guidebook. It deliberately avoids exotic designs and focuses instead on five basic models which appear feasible to implement in real-world settings. Despite this orientation, it must be acknowledged that evaluation is not, and cannot be made simple. Particularly where situational constraints force adoption of statistical rather than experimental controls for extraneous influences, theoretical and computational complexities multiply at an astonishing rate.

It seems likely that some potential users of the guidebook will find certain sections overly technical. On the other hand, those readers who can follow the more difficult portions may find much that seems trivial or unnecessary. Perhaps the best that can be hoped is that a reasonable compromise has been found between the inherent complexity of the total evaluation problem and the need to accomplish meaningful assessments without placing unreasonable demands on the technical expertise of the evaluator.

Organization and Content

The guidebook covers the entire evaluation process from the administrative decisions in selecting an evaluation design to the details of collecting, analyzing, and reporting the data. Many of the details will be of interest primarily to the evaluation specialist and the project

director may skip those sections without detriment. The following paragraphs summarize the topics and indicate the audience for whom each was intended.

The final sections of this chapter describe Evaluation Basics and Preliminary Planning. These sections are quite brief and should be read by project directors as well as those concerned with the details of project evaluation.

Chapter II describes 12 hazards which are commonly encountered in educational evaluation and which may completely invalidate otherwise sound studies. Each of the hazards is named and then described very briefly. Material is then presented discussing why the hazard may invalidate impact assessment. Finally, there is a section on how the hazard can be avoided.

The 12 presentations are not lengthy and should be read by both project directors and evaluators. As a minimum, project directors should read the summary statement of each hazard in order to recognize the practices and to realize that they must be avoided if a valid evaluation is to be done.

Chapter III presents a procedural guide for selecting a suitable evaluation model given the particular set of constraints faced by the project director and evaluator. The entire procedure is presented in decision-tree form (see Figure 1, p.47) with each decision point represented by a question followed by a choice of two alternatives (e.g., Is a comparison group evaluation design feasible?). Each question is discussed on separate pages which describe the implications of the decisions and the alternative courses of action available to the evaluator.

It is strongly recommended that the project director as well as the evaluator read Chapter III. Portions of some of the discussion sections become quite technical and may be skipped by the project director, but it is important that he be familiar with the evaluation options open to him and with the consequences of the decisions he must make.

Chapter IV presents the five evaluation models referred to in Chapter III. There is a brief summary of each model which describes

its general characteristics, its strengths, its weaknesses, and considerations relating to its implementation. The summaries should be carefully read by the project director.

Each summary is followed by several pages of step-by-step instructions for implementing the model except in one instance where the computational procedures were judged too complex for inclusion.

The sections on implementation are intended for use by evaluation specialists and are somewhat technical. It is assumed that the reader will have had at least one college-level course in elementary statistics and will be familiar with and able to compute means, standard deviations, and correlations. No further expertise should be required to follow the implementation procedures, although the underlying concepts and rationales may not always be understood. Consultation with a statistician is advisable for evaluators not familiar with the concepts of covariance and regression if models employing these statistical procedures are selected.

The format of the guidebook is such that the design selection procedure in Chapter III (decision-tree) will automatically lead the reader to only one of the five models described in Chapter IV. He would thus not need to read any of the other model descriptions. Preliminary experience with these chapters, however, suggests that they are interactive, and that reading about the alternative models--particularly the sections dealing with the considerations relevant to implementation--will often lead to a rethinking of the decision made in the design selection procedure. For this reason, at least a superficial reading of all of the model descriptions is recommended before a final model selection is made.

Chapter V deals with the details of data collection and Chapter VI with summarizing and reporting of impact data. These chapters need not be of great concern to project directors although a cursory review of what they contain might facilitate understanding and communication with the project evaluators.

Several appendices are also provided which expand upon issues raised in the body of the guidebook. These appendices, of course, are intended primarily for evaluation specialists and need not concern project directors.

Evaluation Basics

To find out whether students do better in a special project than they would have done without it, the evaluator needs two things: a good measure of how the students performed after their project participation, and an accurate estimate of how they would have done without the project. The difference between these measures provides an index of the project's impact. In order to get a good measure of how students performed, the evaluator must select an appropriate test and ensure that it is administered and scored correctly. Often, the catalog of available tests will not include one with exactly the characteristics desired for assessing a particular project. However, most standardized reading and math tests are sensitive to any significant cognitive growth and should usually prove adequate for assessing the impact of special treatments. Objective-referenced or criterion-referenced tests are also suitable assuming that they have been carefully constructed. Tests and testing are discussed further in Chapter V of this guidebook and in Appendix A.

A more difficult problem lies in estimating how students would have done without the project. In university laboratory studies, the experiences of randomly selected comparison groups are controlled so as to be identical to those of the experimental group in all respects except for the variable of interest. This approach is rarely a viable option in school projects. A variety of substitute approaches are commonly used but all are in varying degrees less satisfactory. The worst of these alternatives are included in Chapter II as "hazards" and make evaluations meaningless. The best are included in Chapter IV with recommendations on when they should be used and explanations of their strengths and limitations.

Chapters V and VI also suggest, as mentioned above, ways of analyzing, interpreting, and reporting results. Details of recommended procedures are included there while characteristics of some widely used commercial reading and mathematical achievement tests are included in Appendix A.

Preliminary Planning

Ideally, the planning of an evaluation should proceed concurrently

with the planning of the project to be evaluated. Obviously, this is not always possible, but it is important to be aware of the fact that some project decisions have important implications for evaluation, and vice-versa. Project-related decisions may, in fact, preclude the possibility of conducting any meaningful kind of evaluation.

One area where close coordination between project and evaluation planning is absolutely essential is that of selecting project participants. Several possibilities exist:

(a) All children comprising a particular group (e.g., all third graders) may be given a special supplementary project
(b) participants may be randomly selected from an identifiable group or population, or
(c) eligibility for participation may depend on the special needs of some members of a larger group (e.g., disadvantaged, gifted).

Each of these alternative participant selection plans fits one or more of the models presented later in this guidebook, but is incompatible with, or places special restrictions on others.

A second area where coordinated planning is required is the matching of evaluation models with test instruments. Criterion-referenced tests can be used with all but the norm-referenced model which requires standardized tests. The norm-referenced model not only requires that standardized tests be used but that the same level of the same test be used for both pre- and posttesting and that the testing be accomplished at exactly prescribed times during the year.

When a project director makes an "executive" decision to use a specific type of test or some particular method of selecting participants he severely limits the number of evaluation models which can be used and may substantially reduce the conclusiveness of his assessment as well. The assumption made in this guidebook is that his first concern will be for conclusive findings. Accordingly, he will wish to consider the feasibility, practicality, and limitations of the more scientifically

sound evaluation models before restricting his choices through hasty decisions about tests or participant selection procedures. In accordance with this orientation, the model selection procedure illustrated in Figur 1 presents the models arranged in order of decreasing rigor from the top to the bottom of the figure so that the evaluation planner can see the consequences of each of his decisions.

Once a model is selected through the decision-tree process, the evaluation planner can read about its strengths and weaknesses and about the conditions and restrictions associated with its use. Careful study of the remaining four models may suggest alternatives that appear more desirable. At that time, he might decide to reject his first choice, re-enter the decision tree, and select another model.

The decision points in the model selection procedure all relate to the manner in which no-treatment, posttest performance expectations are generated. Even where the most rigorous model is selected, however, there are many possibilities for implementation errors which could invalidate the entire evaluation. The next section of this guidebook describes twelve of the most commonly encountered hazards, their consequences, and what should be done to avoid them. These common hazards should be studied carefully before any evaluation is undertaken.

II. COMMON HAZARDS IN EVALUATION

This section describes twelve common hazards in evaluation, the problems they create, and the ways in which the problems can be avoided. The occurrence of any one of the twelve may completely invalidate an otherwise sound evaluation. The hazards include the following:

1. The use of grade-equivalent scores.
2. The use of gain scores.
3. The use of norm-group comparisons with inappropriate test dates.
4. The use of inappropriate levels of tests.
5. The lack of pre- and posttest scores for each project participant.
6. The use of non-comparable treatment and comparison groups.
7. The selection of project participants based on pretest scores.
8. The assembling of a matched comparison group after the project participants are selected.
9. The careless administration or scoring of tests.
10. The assumption that an achievement gain is due to the treatment when, in reality, it is due to some other factor.
11. The use of non-comparable pretests and posttests.
12. The use of inappropriate formulas to estimate posttest scores.

Although subsequent sections of this guidebook refer back to specific hazards, it is strongly recommended that the reader study all of the hazards before going on to other material.

Hazard 1

The use of grade-equivalent scores.

Grade-equivalent scores provide an insensitive, and in some instances systematically distorted, assessment of a project's impact.

Why is this a hazard?

There are three serious problems with grade-equivalent scores:

A. The concept of a "grade-equivalent" score is misleading. For example: a grade-equivalent score of seven attained by a fifth grader on a math test does not mean that he knows sixth- and seventh-grade math. It is more accurate to say that he can do fifth-grade math as well as an average seventh grader can do fifth-grade math although even this representation is not strictly accurate. It is quite possible, in fact, that when the test was normed no seventh graders ever took the level of the test intended for use in the fifth grade. In such cases, the seventh-grade grade-equivalent scores reported in the test manual are simply statistical projections and tell us little about how seventh graders would have actually scored if they had taken the fifth-grade test.

B. Grade-equivalent scores do not comprise an equal-interval scale. That is, a grade-equivalent score of two is not in any sense "half" of a score of four. For this reason, "average" grade-equivalent scores are not consistent with averages computed from more appropriate kinds of scores and are not interpretable.

C. The normative data for many commercial tests are collected during one short interval of the school year, often in February or March. In order to establish norms for fall and spring, a smooth curve is drawn connecting the points which represent actual data. Unfortunately, there is substantial evidence that learning does not proceed uniformly over

the calendar year. One factor that contributes to the irregular learning pattern is the effect of reduced gains or even forgetting during the summer months. As a result, the current procedures used to generate grade-equivalent scores tend to make them systematically too low in the fall and too high in the spring. Fall to spring gains thus appear to make a project look unusually effective when actually the gains were exactly the same as would be expected under normal classroom conditions. Even where using grade-equivalent scores does not introduce systematic biases (as would be the case if there were a 12-month pre-to-posttest interval), the curve-fitting procedures used to generate such scores introduce errors large enough to invalidate any evaluation.

How can the hazard be avoided?

There is never any technically sound reason for using grade-equivalent scores in evaluating projects and they should be avoided. Standardized tests can still be administered. However, raw scores should be converted to standard scores instead of grade-equivalent scores before summary statistics are computed. Mean pretest and posttest standard scores should then be converted to their percentile equivalents. Finally, pre-to-posttest gains can be compared against expectations derived from the national norms, but only if the tests were administered at appropriate times (see Hazard 3). Further discussions of the problems created by using grade-equivalent scores for evaluation purposes can be found in Tallmadge and Horst (1974, Appendix E).

Hazard 2

The use of gain scores.

There are several kinds of gain scores and they are generally used in an attempt to adjust for initial differences between treatment and comparison groups in conventional experimental designs. Where such differences are real, they cannot be adequately adjusted. Where they are the result of random sampling fluctuations, "raw" gain scores overcorrect for between-group differences and "residual" gain scores are likely to undercorrect.

Why is this a hazard?

The most commonly encountered type of gain score is the "raw" gain score which is simply the posttest score minus the pretest score. The term "raw" does not refer to the type of pretest or posttest scores (raw, standard, percentile, etc.) used to determine the gain but to the gain itself. The category of raw gain scores thus includes grade-equivalent gains.

If differences between treatment and comparison groups are random, (i.e., the two groups may be regarded as random samples from a single population) then raw gain scores overcorrect for pretest differences by excessively inflating the posttest performance measure of the initially inferior group. Analysis of covariance provides an appropriate means of adjusting for random pre-treatment differences between groups.

In certain theoretical situations where differences between treatment and comparison groups are real (i.e., the groups are samples from different populations) gain scores may represent the best method for equating the groups. In real-world educational evaluations, however,

factors such as differential growth rates, different test score reliabilities as a function of achievement level, different reliabilities on pre and post measures, and test floor effects all work to invalidate this type of adjustment (Campbell, 1974). The authors reject statistical techniques for equating truly non-compatable groups in conventional experimental designs. Such groups only permit defensible conclusions of effectiveness in those rare instances when an initially inferior treatment group outperforms the initially superior comparison group on the posttest.

A "residual" gain score is not a gain score at all. It is the difference between an actual posttest score and an estimated posttest score where the estimate has been derived from the pretest scores using regression techniques. Whenever there is a pretest difference between treatment and comparison group means, residual gain scores systematically undercorrect. The amount of undercorrection is directly proportional to the size of the between-group difference.

Tallmadge and Horst (1974, pp. 36-37) presents a further discussion on gain score problems.

How can the hazard be avoided?

Gain scores should never be used. Where pretest scores are equal for treatment and comparison groups, there is, of course, no need for the kind of adjustment gain scores are supposed to provide. Where between-group differences result from random sampling fluctuations, covariance analysis is the appropriate technique to use. Where the differences are real, and the groups are truly non-comparable, there is no adequate technique for equating them and conventional comparison group evaluation models should not be used. Appropriate alternative models are recommended in Chapter III.

Hazard 3

The use of norm-group comparisons with inappropriate test dates.

Administration of tests on dates which do not correspond to the date when the actual normative data were collected invalidates norm-referenced comparisons.

Why is this a hazard?

When comparison groups are available, few evaluators would even consider testing treatment and comparison students more than a few days apart. When a norm group is used for comparison, this issue appears to be given little thought. The problem stems from two misleading practices followed by test publishers. First, interpolation or extrapolation processes are used to "create" norms for periods when no "real" normative data were collected. Thus, most publishers provide norms for fall, winter, and spring even though data were collected at only one or possibly two of these points. Projected norms are generally based on the assumption of linear cognitive growth over each month of the nine months of the school year with one additional month's gain over the three summer months. There is no evidence to support this assumption and the created norms are likely to be far enough off to distort the impact of special instructional projects. Norms based on projected estimates should never be used for evaluation purposes.

The second practice is the suggestion, implicit in most norms tables, that the norms are valid over a three- or even a four-month period. For this to be the case, children would have to learn nothing over the entire period, then show a large gain overnight at the end of the period and so on. This assumption is clearly absurd. If the norms are correct somewhere in the middle of the time period, they will be systematically too

low at the beginning and systematically too high at the end of the period. The errors involved are quite large and can give a severely distorted picture of project impact. (See Tallmadge & Horst, 1974, Appendix D, p. 67.)

How can the hazard be avoided?

It is absolutely essential to test children in the treatment condition within a week or so of the dates on which the norm groups were tested. Tests which provide normative data for only one point in the year should not be used for norm-referenced evaluation of fall-to-spring gains. Instead it is better to select a test with normative data in both fall and spring even though the choice of tests is then limited. Basically, it is never advisable to extrapolate or interpolate very far from observed normative data.

Hazard 4

The use of inappropriate levels of tests.

If most of the pupils tested are getting nearly all, or hardly any of the test items correct, the level of the test is inappropriate for assessing their cognitive achievement status. Measurement under these conditions is both unreliable and invalid. Ideally, the pupils tested should score in the middle of the range of possible raw scores.

Why is this a hazard?

The major standardized achievement tests are divided into several levels which cover different grades or grade bands. Each level is an individual test appropriate for only two or three grades. In the case of projects aimed at slow or fast learners, the test level nominally designated for their grade is likely to be too difficult (pupils will encounter the test "floor") or too easy (pupils will encounter the test "ceiling") and would not provide a reliable and valid measure of achievement. Ceiling and floor effects may cause similar distortions in evaluations using criterion-referenced tests.

How can the hazard be avoided?

Test levels should be selected on the basis of the achievement levels of the pupils, not on the basis of their grade in school. Usually, one level above or below that nominally recommended for a particular grade will be sufficient to avoid ceiling and floor effects, but no firm recommendation can be made as difficulty levels and ranges of coverage vary greatly from instrument to instrument.

Using test levels other than those nominally recommended for

particular grade levels is likely to mean that norms tables for the grades tested are not included in the test manuals. This is unfortunate since it is clearly not meaningful to assess either status or growth through comparisons with children at a different grade level. The status of a sixth grader should be assessed using sixth-grade norms even if he is tested with a fourth-grade test. If a comparison group is available, there is no problem because growth is assessed with reference to the comparison pupils--not to the norms. With norm-referenced evaluation models, however, there may be a problem. Fortunately, most major test publishers have interlocked their test levels by providing overlapping grade-level coverage. This practice has enabled the development of score equivalencies between adjacent test levels so that it is possible to predict quite accurately from a pupil's score on one test level how he would have scored on the next higher or lower level.

From the between-level score equivalencies, it is common practice to develop a single score scale spanning all test levels so that raw scores from any level can be converted to scores on this scale (scores of this type are often called scale, standard, or expanded standard scores). Scale scores can be referenced to any set of normative data. Thus, scores of sixth graders tested with a fourth-grade test can be converted to sixth-grade percentiles and it is not necessary to use a test which is likely to be too easy or too difficult for the particular children being tested. While there are generally some measurement errors which result from imperfect interlocking, typically they are smaller than those which result from encountering test ceilings or floors.

Whatever level of a test is selected for use, that same test level should be used for both pre- and posttesting (see Hazard 10).

Hazard 5

The lack of pre- and posttest scores for each treatment participant.

Analyses of project impact should be based only on those participants with both pre- and posttest scores. Interpretation of these data, however, should take into consideration the characteristics of pupils who dropped out, entered late, or graduated from the project.

Why is this a hazard?

In most projects, the group that is ultimately posttested is not composed of exactly the same students as the pretest group due to drop-outs and new students during the school year. Therefore, pre- and post-test mean scores are not strictly comparable. In particular, it often seems that the dropouts from a special program are among the slowest students. Eliminating their low scores from the posttest may raise the mean posttest score considerably. On the other hand, some projects may return successful students to their regular classrooms, thus lowering the mean posttest score for the remaining group. It is not uncommon to find evaluation reports which include posttest scores for fewer than half of the reported project participants and any conclusions in such reports are usually meaningless.

How can the hazard be avoided?

It is not possible to prevent students from dropping out of or entering a project after it has begun. Still, it is essential to base any conclusions about the impact of the project on the data from students who have both pre- and posttest scores, but even this is not enough. The pretest score distribution for all dropouts must be examined to see if it differs from that of the non-dropouts. Further, if the number of dropouts

is large, at least a brief investigation of the reasons for dropping out is required. Sometimes a project is targeted at certain children, and the dropouts may be either students who succeeded and returned to their regular classes leaving the unsuccessful students to be posttested, or they may be exactly the students for whom the program was intended, but who failed and left.

In short, every effort must be made to obtain pre- and posttest scores for each project participant. Pretest-posttest comparisons must be based on those students for whom both scores are available. Data from students having only pretest or only posttest scores must be carefully examined to see if they differ in some systematic respect from the data of students having both pre- and posttest scores. A description of any of these differences should be included in the project evaulation. Analysis of pre- and posttest scores is discussed further in Chapter VI.

Hazard 6

The use of non-comparable treatment and comparison groups.

In conventional experimental designs, treatment and comparison groups must be comparable in all relevant variables before the treatment begins. Groups which differ in terms of pretest scores are an obvious source of bias. Other, more subtle factors such as differences in age, sex, race, or socioeconomic status can also exert strong biasing influences and must be avoided. In such designs, there is no way in which a non-comparable comparison group can provide an accurate estimate of how well the treatment group would have done without the treatment.

Why is this a hazard?

Students in a special program may do better or worse than comparison groups simply because they were different to start with. One of the most common cases occurs when students who volunteer are put in the special program while the rest serve as a comparison group. Even given equal pretest scores, it is likely that the volunteers are a more enthusiastic group and will learn more. This type of rather subtle difference is often overlooked. Of course, any obvious differences between treatment and comparison groups may also affect evaluation results and such variables as socioeconomic status, age, sex, racial and ethnic composition, and school size and setting should be carefully checked for comparability.

The problem is even more serious when norm-based comparisons are used. Volunteering or other selection procedures may result in a treatment group that is quite different from the norm-group students who got equal scores at pretest time.

The net result in either case is that the comparison group provides

an inaccurate estimate of what project participants would have learned without the project treatment. Theoretically, the estimate may be either too high or too low. However, typical selection strategies usually lead to superior treatment groups.

How can the hazard be avoided?

Students should be assigned to treatment and comparison groups on a random basis or in such a way that a nonrandom assignment is random in effect (Lord, 1967, p. 38). Essentially, this means that the two groups must be similar along all educationally relevant dimensions, unless the evaluation model specifically allows for group selection on the basis of pretest scores. This hazard and the steps to avoid it are closely related to the previously discussed Hazard 2.

Hazard 7

The selection of project participants based on pretest scores.

When students are selected for project participation based on their obtaining relatively high or relatively low scores on some test, use of those scores as pretest measures invalidates any kind of norm-referenced evaluation.

Why is this a hazard?

This error has been so widely discussed and well documented that most evaluators are aware of the problem. Unfortunately, for various reasons it is still encountered. The error results from testing a large group of students, selecting the lowest (or highest) ones for a special program, and then treating the selection scores as pretest scores. This practice results in systematic distortions on pre- to posttest gains.

It is well known that if the low scoring students are retested on the same or a comparable test, they will score higher on the average, while an initially high scoring group will score lower. This phenomenon is called "regression toward the mean," or simply "statistical regression," and is discussed in virtually all texts on experimental design. The result is that low scoring groups appear to learn more from a special program than they actually do, while gains in special programs for high scoring students may be obscured.

Statistical regression presents no problem for the special and general regression models presented in Chapter IV. Evaluations employing comparison groups may or may not be affected depending on whether the regression effect operates differently on the two groups. Hazard 8 treats a closely related situation in which the comparison group is selected on the basis of pretest scores. Regression artifacts invalidate any kind of norm-referenced evaluation.

How can the hazard be avoided?

Corrections for the regression effect are possible in theory, but in practice the necessary data are not usually available. Thus, it is safer to avoid the problem by not using the pretest to select project participants except for those regression models which specifically require this approach. (See also Step 7, p. 23 of Tallmadge & Horst, 1974.)

Hazard 8

The assembling of a matched comparison group after the project participants are selected.

Finding "matches" for treatment participants in some other group is a fundamentally unsound practice. Unless they and the treatment pupils are equally representative of the groups from which they are drawn, statistical regression will act differentially on the two groups and artificially inflate the apparent gains of one group with respect to the other.

Why is this a hazard?

It may be very useful to have a comparison group made up of students carefully matched to the treatment students, but unless the proper procedures for selection are followed, comparisons between the two groups may be completely misleading. The common practice of selecting students for the treatment, then trying to find a non-treatment student to match each treatment student is a serious evaluation error. If, for example, a project is set up for the most underachieving children in a disadvantaged school, it may be possible to construct a "matching" comparison group by finding children with equally low pretest scores in less disadvantaged schools. In this situation, the comparison students would be farther below the means of their own schools than the treatment children and their posttest scores would show a greater regression toward the mean. This regression artifact would thus inflate the apparent gains of the comparison group with respect to the treatment group and might obscure a real project impact.

How can the hazard be avoided?

The correct procedure for establishing matched comparison groups

is to do the matching first and then assign members of each pair randomly to the treatment or the comparison group. That is, a large group of students, all eligible to be in the project, must be available. The first step is to divide the group into matched pairs based on test scores, ethnic background, sex, etc., so that the two members of each pair are as similar as possible. Then, after the matching process is complete, some random procedure such as flipping a coin is used to decide which student goes into the treatment and which into the comparison group. Where this approach is impossible, models which do not require matched groups should be selected. (See Chapter III.)

Hazard 9

The careless administration or scoring of tests.

Testing must be accomplished with scrupulous attention to detail. For most evaluation models, the primary requirement is that treatment and comparison groups be tested in exactly the same way. The norm-referenced evaluation model further requires that procedures outlined by test publishers be followed precisely.

Why is this a hazard?

Problems arise if tests are administered or scored in an inconsistent and careless manner. If there are differences in the ways in which the treatment students and the comparison students are tested or if there are differences in the procedures, conditions, and scoring at pretest and posttest times, then it is impossible for the resulting data to accurately reflect project impact. No amount of careful statistical analysis can later overcome these problems.

How can the hazard be avoided?

a) Test procedures must be orderly and accurate if scores are to be meaningful.

b) The treatment students must be tested and scored in exactly the same way as comparison students.

c) The procedures, conditions, and scoring methods during posttesting must be exactly the same as during pretesting.

Properly trained personnel decrease the probability of disorderly or inaccurate testing procedures but problems may be introduced by local conditions and student attitudes. Students may not understand what is

expected of them, or in extreme cases, they may become unruly and make no serious effort to answer test questions. Problems which occur due to carelessness include failing to get the right name on each answer sheet, using the wrong answer key or conversion tables, and making mistakes in copying scores onto data sheets.

The second issue, comparability between the testing situations of the treatment and comparison groups, can and should be dealt with in a straightforward manner in comparison-group designs. In these cases, identical procedures, even the use of a single tester, are possible. In the more common situations in which norm-group comparisons are made, the instructions accompanying the test must be followed exactly.

The third issue, comparability between pre- and posttesting situations, requires the same attention to procedures as the other issues. The real problem is often the pressure on teachers to show achievement gains which may lead them, intentionally or unintentionally, to be stricter in enforcing time limits and avoiding helpful hints on the pretest than when administering the posttest. This type of problem can be minimized by having an independent, external evaluator administer the tests or by having teachers within a school exchange classrooms so that each tests and scores another teacher's students.

Chapter V is devoted entirely to the details of obtaining accurate, meaningful data.

Hazard 10

The assumption that an achievement gain is due to the treatment when, in reality, it is due to some other factor.

Other possible explanations always exist for observed gains. The plausibility of these alternative explanations should be carefully examined before gains are attributed to project impact.

Why is this a hazard?

Sometimes project participants learn substantially more than would have been expected, but the project, per se, is not responsible. Instead, the gains could be a result of the Hawthorne effect (Whitehead, 1938) in which special project participants do well simply because they are getting special treatment. The nature of the treatment may not necessarily be important. An opposite result may follow from a John Henry effect (Saretsky, 1972). In this case, comparison-group students work extra hard to prove that they are just as good as project students.

Other likely causes of misleading gains are unrecognized "treatments" which have nothing to do with the project. Most school systems are in a constant state of flux with multiple changes every year. Changes in school programs, personnel, facilities, class sizes, community characteristics--any or all of these factors can affect student performance. Also, the true source of achievement gains is sometimes improperly identified because children are involved in more than one treatment. Under these conditions it is impossible to determine causality in an unambiguous manner.

How can the hazard be avoided?

When a carefully implemented evaluation reveals significant cognitive achievement gains, it should not be immediately assumed that the

gains are solely the result of the special treatment. A variety of other factors exist which may lead to the obtained results. Each plausible rival hypothesis should be examined and, where the evidence permits, eliminated as a likely explanation. A discussion of the remaining factors and the relative likelihood of each as a contributor to the gains should be included in the evaluation.* In succeeding years with a continuing project, some of these competing explanations might be controlled and eliminated.

Hazard 11

The use of non-comparable pretest and posttest.

It is almost always a good idea to use the same level of the same test for both pre- and posttesting. In norm-referenced evaluations, it is usually essential.

Why is this a hazard?

The situation in which pretests differ from posttests is frequently encountered in evaluation reports. Usually it occurs because there is a district-wide change in testing policy during the evaluation period in an attempt to find a more appropriate test for all district evaluations. The disruption of evaluations of ongoing projects is unavoidable, and may be completely beyond the control of the project evaluator. It may also, however, severely limit the usefulness of the evaluation and should be avoided if at all possible. The use of the same level of a test for both pre- and posttesting is also strongly advised. Some tests have interlocked levels so that scores from one test level can be converted into another. However, these conversion tables reflect a certain degree of measurement error as a result of curve fitting, rounding, and successive transformations. It is clearly preferable to use just one level of the test.

In a comparison-group design, the fact that the posttest differs from the pretest may not be a critical problem. So long as pre- and posttests are reasonably correlated, as will be true among the major commercial tests, the comparison-group students make reasonably convincing conclusions possible. However, in the more common norm-referenced designs, there is no completely adequate way to compare pretest scores on one test

with posttest scores on a completely different test. Since each test is normed on a different group of students, this amounts to using one comparison group for the pretest, and a second comparison group for the posttest.

How can the hazard be avoided?

To insure comparability between the pre- and posttests in norm-referenced evaluations, the only real solution is to administer the same level of the same test on both occasions. When that option is not available, it still may be possible, in some instances, to approximate it through the use of conversion tables provided in the Anchor Test Study (Loret, Seder, Bianchini & Vale, 1974). The Anchor Test Study provides tables which may be used to convert scores on one test to their equivalents on each of the other tests in the study. Conversion errors are reported to be low, so in theory the procedure is sound, but, in any case, it applies only to the eight most commonly used reading tests covered by the study, and only to grades 4, 5, and 6.

In comparison group evaluations, switching from one standardized test to another is acceptable if both tests meet the requirements of this guidebook. The result is usually to lower pretest-posttest correlations and correspondingly to lower precision of the evaluation. Switching to an entirely dissimilar test is to be strongly discouraged.

Hazard 12

The use of inappropriate formulas to estimate posttest scores.

Under certain circumstances, it makes sense to expect that a pupil will maintain his relative status with respect to national norms from pre- to posttest if he does not participate in a special project. However, many methods have been devised for calculating performance level expectations which rest on clearly untenable assumptions. These methods of estimating performance levels should never be used.

Why is this a hazard?

Many projects use an unrealistic theoretical model or formula to calculate "expected" posttest scores from IQ or other pretest scores. If students do better than the calculated expectation, the project is considered a success. Estimated posttest scores are often based on average grade-equivalent scores. For example, a student who has gained 0.7 years per year, on the average, since beginning school is presumed to continue at the same rate unless a special program increases his rate. There are many problems with such an estimate, but the major one is in the use of grade-equivalent scores (see Hazard 1). The student who averaged 0.7 years per year over several years will usually appear to gain more than that if measured from fall to spring, giving a misleading impression of improvement.

Most IQ-based estimates are both inaccurate and logically unreasonable. For example, the Bond-Tinker formula (Della-Piana, 1968, p. 41) is often used to compute an "expected" reading level, i.e.,

$$\text{Expected reading level} = \left[\frac{\text{IQ}}{100} \right] \times [\text{ No. of years in school }] + 1.$$

For a student with an IQ score of 85 (approximately one standard deviation below the mean) at grade level 7.1 (6.1 years of school completed):

$$\text{Expected reading level} = (.85) \times (6.1) + 1 = 6.2$$

So the formula says he should be reading at the sixth-grade level. But since his IQ is supposed to be "mental age" divided by "chronological age," his mental age would be given by:

$$MA = (IQ) \times (CA)$$

Assuming the seventh-grader is twelve years old:

$$MA = (.85) \times (12) = 10 \text{ years}$$

We now have a twelve-year old student with a mental age of ten years who is expected to read as well as an average sixth grader (11 years old). This is certainly inconsistent, but even worse, it is incorrect. According to normative data from the Gates-MacGinities reading test, a seventh-grade student one standard deviation below the mean is reading at the fourth-grade level.

Because of these and many other theoretical and practical problems, the underlying concepts of the intelligence quotient have been abandoned by informed measurement specialists (Cronbach, 1970, p. 216; Tyler, 1972, p. 177). While the commercial instruments which have been designed as "IQ tests" may have a variety of practical uses, they are not, in general, the best available predictors of specific school skills, and IQ scores are not recommended for any purpose in evaluating the effects of special projects.

How can the hazard be avoided?

In norm-referenced evaluation models, posttest scores can be estimated by referring to national norms. When comparison groups are used, the actual posttest scores of these groups, or a regression equation estimating the posttest scores, provide the proper basis for evaluating treatment effects.

III. A PROCEDURAL GUIDE FOR MODEL SELECTION

This section presents a procedural guide for selecting an evaluation model. By answering a series of questions relating to the real-world constraints under which the evaluation will be conducted, the reader is led to one of the five evaluation models presented in the following chapter.

Figure 1 on page 47 summarizes the seven-step decision tree in flow-diagram form. Each step is discussed separately on the pages preceding Figure 1. (This page arrangement is intended to facilitate reference to the fold-out figure.) For each step, the decision question is presented along with two answer alternatives. A "comment" section is also included which explains the issue in question and the implications of choosing each alternative course of action.

The specific path to be followed through the decision tree depends on the answers the reader makes to each of the seven questions, and instructions on how to proceed are provided for each answer alternative. The reader should first read through the chapter and then make a selection by skipping from page to page in accordance with these instructions.

Figure 1 also shows the five evaluation models which are discussed in Chapter IV. They are arranged in decreasing order of scientific rigor, with those at the top of the page enabling the evaluator to draw substantially more conclusive inferences about project impact than those at the bottom. On the other hand, the feasibility of implementation is expected to operate in exactly the opposite direction so that the less rigorous models will be much easier to use. While the more rigorous models are certainly to be preferred, any one of the five will yield believable results if carefully implemented.

Question 1

Do practical considerations (policy, availability, cost, time) permit you to select an evaluation design which makes use of a local comparison group?

Yes Proceed to Question 2

No Go to Model 5, page 72

Comment

In order to measure the impact of any special instructional treatment, it is essential to have some estimate of how the participants would have fared under normal or non-treatment conditions. Since, presumably, the non-treatment condition consists of participation in a regular school curriculum, some gains would clearly be expected even without the special project. The problem is to obtain a good estimate of how large the pupils' gains would have been under such conditions and subtract this estimate from the gains they actually obtained in the special project. The difference is the incremental gain which can be attributed to project participation.

There are two kinds of local comparison groups which can provide adequate estimates of non-treatment expectations: (a) a conventional comparison group which is like the treatment group in all educationally relevant respects, and (b) a comparison group which results from splitting a pre-existing intact group into treatment and comparison subgroups at some pretest cutoff score.

The best method of estimating non-treatment posttest scores is to find a group of pupils exactly like the project children and to treat them in exactly the same way with the single exception of withholding the special treatment from them. Their posttest scores will then constitute the best possible estimate of how well the treatment group would have done without the treatment.

It is often not possible to obtain a sample of exactly comparable pupils to serve as a comparison group. Under appropriate conditions, however, groups which are not strictly comparable can be used for estimating non-treatment performance. Model 3, in fact, divides a class or other pre-existing group into treatment and control subgroups at some pretest cutoff score so that all pupils above the cutoff go into one group while all pupils below it go into the other group.

The issue of comparison group suitability and the implications which the type of group has for selecting an evaluation design are addressed in subsequent Questions. If either type of local comparison group is available, proceed to Question 2.

Where no local comparison group is available, the evaluation must depend on comparisons between treatment-student scores and national norm-group data collected by the publishers of standardized tests. This procedure is explained in Model 5, page 72.

Question 2

Will pre-existing, intact groups or individual pupils be assigned to treatment and comparison conditions?

Groups Skip to Question 5

Pupils Proceed to Question 3

Comment

The most commonly encountered type of intact group is a classroom, a school, or a grade level within a school. Assignment by groups would mean that one third-grade classroom was assigned to the treatment condition and another to the comparison condition--or that all third graders in one school comprised the treatment group while all third graders in another school constituted the comparison group. Third graders from one school who were in the lowest quartile of the national distribution in reading could also be considered a pre-existing, intact group if they were compared against similar children from another school. In all of these cases, the condition to which the pupils were assigned was determined entirely by their group membership without regard to any characteristics of the individuals.

On the other hand, if all third graders were listed alphabetically and alternately assigned to treatment and comparison conditions, we would say that assignment was by individual pupil. Another similar example would entail the pairing of children on the basis of their pretest scores with subsequent assignment of one member of each pair to the treatment group and the other to the comparison group.

A quite different kind of assignment, but one still considered assignment by pupils, involves the assignment of pupils who score below some selected cutoff point on a test to the treatment group and those scoring above that point to the

comparison group. In this case, some members of an intact group were assigned to the treatment condition and others to the comparison condition but it should be clear that assignment to conditions was based on considerations relating to individual characteristics and not group membership.

Assignment by pupil is generally preferred over assignment by group as this method offers greater control over potentially biasing factors. The use of pre-existing groups is a viable alternative only where the groups are similar in all relevant respects to groups which would have resulted from assignment by pupil.

Question 3

Is it possible to assign pupils randomly to treatment and comparison groups, or will group membership be determined by need?

Randomly Proceed to Question 4
By Need Skip to Question 6

Comment

Random assignment implies that each child in a single "pool" or group has an equal chance of being assigned to the comparison or to the treatment group. One way to accomplish random assignment would be to place the names of all the children in a hat and then draw them out one at a time assigning every other child to the treatment group. There are other techniques which are equally suitable but the decision as to whether a child is assigned to one group or the other must be left purely to chance. Group assignment based on teacher preferences, children volunteering, or similar human actions are not random. To consider them so may be seriously misleading (see Hazard 6, p. 19).

The assumption of random assignment underlies most statistical tests. A statistically significant *t* or *F* test means simply that the observed difference between groups was larger than would normally be expected to result from random assignment. This, in turn, implies that if assignment was random the observed difference was probably due to the treatment. If assignment was not random, however, a "statistically significant difference," by itself, is generally meaningless.

Special projects are most often designed to serve particular segments of the population (e.g., disadvantaged, gifted, bilingual). Under certain circumstances, children in such categories can be selected from a heterogeneous group and

given a special treatment while the remaining children serve as a useful and valid comparison group. Questions 6 and 7 describe the conditions and procedures for implementing evaluation models of this type.

Question 4

Is it possible to match pupils on the basis of pretest scores before randomly assigning one member of each pair to the treatment group and the other to the comparison group?

Yes Go to Model 1, page 49

No Go to Model 2, page 54

Comment

Random assignment usually results in some small differences between groups in terms of pretest performance. At least some of this difference can be expected to carry over to posttest performance. For this reason, it is desirable to remove these differences, however small, either by pre-assignment matching (see Model 1, page 49) or by statistical manipulation after the fact using analysis of covariance (see Model 2, page 54). Pre-assignment matching is the preferred technique if feasible and has the additional advantage of minimizing computational complexity--a significant drawback of covariance analysis techniques.

Matching must be accomplished before pupils are assigned to groups. The correct procedure is to identify pairs of students having equal or essentially equal scores on some test known to correlate highly with the post-treatment measure. One member of each pair is then assigned to either the treatment or the comparison group based on the outcome of some random event such as the flip of a coin. The remaining member of each pair is assigned to the other group.

One of the most common errors in educational evaluation is that of matching after assignment. If, for example, there are two pre-existing groups, it is common to administer the treatment to one of them while selecting pupils with matching pretest scores from the other to serve as a comparison group.

Although common, this procedure is fundamentally unsound and introduces systematic biases into the data. Unless matching can be accomplished prior to assignment it should not be done at all. (See Hazard 8, page 23.)

Question 5

Where a pre-existing comparison group is available, is it sufficiently similar to the treatment group so that the assignment of pupils to groups can be considered "random in effect?"

Yes Go to Model 2, page 54
No Skip to Question 7

Comment

As discussed in the Comment accompanying Question 3, statistical tests of the difference between the means of two groups generally rest on the assumption that group membership was determined through random assignment processes. It is possible, of course, for no educationally relevant differences to exist between two classrooms of third graders in a particular school, or between grade-level peers in two schools in a district. Under these circumstances, the groups are virtually identical to groups which would have resulted from random assignment and their composition may be considered random in effect (Lord, 1967, p. 38).

Where pre-existing, intact groups are used as treatment and comparison groups, it is not appropriate to assume that they are adequately similar. This possibility must be investigated empirically and the onus of proof is on the evaluator. Ideally, the process by which students were assigned to the two groups should have been effectively random. At the very least, the two groups must not be significantly different in terms of pretest scores. They must also be comparable in terms of socioeconomic status, age, sex, and racial composition. School size and setting (urban - rural) as well as neighborhood should also be comparable. Even when these factors are equated, serious

biases are possible. Such biases are introduced when teacher or student participation is voluntary or when the choice as to which group will be the treatment group and which the comparison group is made by principals or teachers. This guidebook discourages any use of local comparison groups which are clearly dissimilar to the treatment group (see Hazard 6).

Question 6

Is assignment to the treatment or comparison group based on a cutoff value on some pre-treatment measure or combination of measures?

Yes Go to Model 3, page 59

No Proceed to Question 7

Comment

Where the memberships of the treatment and comparison groups are neither random nor random in effect, so called "true" experimental designs can no longer be used. Under these circumstances "quasi-experimental" evaluation models must be employed.

There are two quasi-experimental evaluation models (the Special Regression Models) which can provide acceptably conclusive evidence regarding treatment impact in situations where the assignment of pupils to treatment and comparison groups is based on need rather than randomization. Both of these models, however, require the establishment of a cutoff score above which all pupils are assigned to one group and below which all pupils are assigned to the other. Numerical ratings by teachers, classroom grades, and standardized achievement test scores may be used singly or in any desired combination, but there must be a single cutoff score.

Other models exist which do not require assignment to treatment and control conditions based on a single cutoff score. As design requirements of this type are relaxed, however, additional assumptions must be made in order to attribute the cause of observed between-group differences to treatment influences, and credibility is thus diminished. These models are treated in Question 7.

Question 7

Is there a pre-existing comparison group whose performance on the pretest measure is superior to the performance of the treatment group?

Yes Go to Model 4, page 70

No Go to Model 5, page 72

Comment

Quasi-experimental designs all rest on sets of assumptions having varying degrees of plausibility. One such assumption which is relevant here and appears "safe" is that a group which is initially superior to another group in cognitive development will continue to grow at a rate equal to or greater than that of the initially inferior group, other things being equal. If, under these circumstances, the initially inferior group outperforms the initially superior group after participation in a special instructional treatment, it is probably safe to conclude that the treatment was effective. On the other hand, if an initially inferior group receives the treatment but fails to surpass the comparison group on the posttest (a typical situation) it is difficult to draw conclusions with confidence. Under certain conditions regression models not requiring single cutoff scores may be applicable (see Model 4). Finally, if the treatment was administered to the initially superior group and its posttest performance remained superior to the comparison group, it would be difficult to decide whether the superior posttest performance resulted from the treatment or simply from the inherent superiority of the treatment group

If the only available comparison group scores significantly lower on the pretest than the treatment group, the information obtainable from it is usually not worth the time and

expense to collect. A norm-referenced evaluation model will probably be more useful and will certainly be less costly (see Model 5).

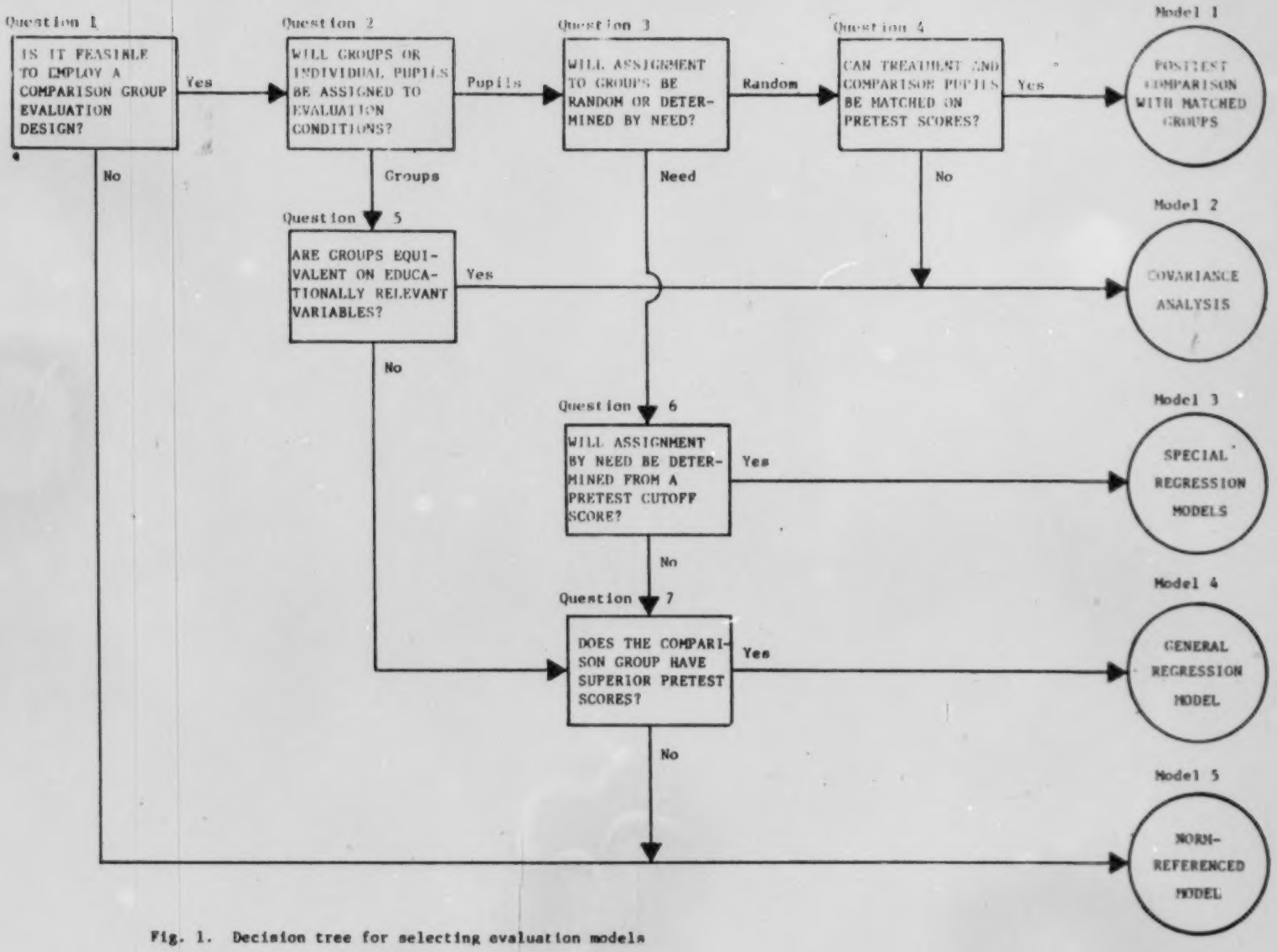

Fig. 1. Decision tree for selecting evaluation models

IV. EVALUATION MODELS

This section of the guidebook provides descriptive information about five evaluation models suitable for use in assessing the cognitive benefits resulting from local school projects. They are not necessarily the only models suitable for this purpose but they are recommended as the most convincing models that can be feasibly implemented given the constraints of operating school systems.

The five models are:

1. Posttest Comparison with Matched Groups (p.49)
2. Analysis of Covariance (p. 54)
3. Special Regression (p. 59)
4. Generalized Regression (p. 70)
5. Norm-referenced (p. 72)

Each of these models is described in terms of general characteristics, strengths and weaknesses, and considerations related to its implementation. Except where computational procedures are excessively complex and require the skills of a sophisticated statistician (the Generalized Regression Model), step-by-step procedures are provided for using each of the models. References to sources of more detailed information are also included.

Each of the evaluation models in this section has specific analysis requirements. However, several preliminary steps are useful with any evaluation model. These preliminary steps are discussed in Chapter VI.

Model 1

Posttest Comparison with Matched Groups

Summary

General Characteristics. This model requires that children be paired in terms of pretest measures and that one member of each pair be randomly assigned to the treatment group and the other to the comparison group.

Strengths. The matched groups evaluation model provides what is theoretically the most accurate estimate of how the treatment group would have done had they not received the special instructional treatment. This high degree of accuracy is due to the fact that the comparison group is constructed so as to be virtually identical to the treatment group at pretest time. Thus, if the experiences of the two groups are the same between pre- and posttest with the single exception of exposure to the treatment, the comparison groups should achieve posttest scores which are essentially the same as those which would have been achieved by the treatment group had its members not received the treatment.

Weaknesses. The manner of assigning pupils to treatment and control groups employed in this model may produce a greater awareness of group membership than other, less obtrusive assignment procedures. Children in the comparison group may realize that their group is not inherently different from the treatment group, yet, for some reason, the other group of children is receiving special attention. This increased awareness of group membership may magnify such spurious influences as the Hawthorne effect in the treatment group or the John Henry effect in the comparison group (see Hazard 10).

Implementation Considerations. This evaluation model allows a wide choice of test instruments and testing times. However, if, as would be recommended, a norm-referenced comparison is also employed, the choice of tests and testing times becomes more restricted (see the Norm-referenced Model, page 72).

According to this model, children in the treatment and comparison groups are matched on the basis of their pretest scores and possibly other educationally relevant variables as well. At posttest time, it is important to use an instrument that measures the same skills as the pretest. If children in a math project are matched on the basis of pretest reading test scores or IQ and then are given an arithmetic posttest, the increase in precision which can be achieved through matching may be substantially reduced. This is because the precision gained is proportional to the correlation between pre- and posttest scores.

In order to implement the model, it must be possible to (a) assemble a group of children large enough in number to form both a treatment and a comparison group, (b) pretest the entire group, (c) pair children on the basis of their pretest scores, and (d) randomly assign one member of each pair to the treatment group and the other to the comparison group. If eligibility for participation in the treatment group is based on some special educational need, this procedure is clearly not feasible and one of the other evaluation models should be implemented.

Implementation Procedures.

Step One: Identify a group of potential participants large enough in number to form both a treatment and a comparison group.

Step Two: Administer the pretest to the entire group with an instrument known to correlate highly with the measure selected for use as the posttest.

Step Three: Score the pretest. Using their raw scores, identify pairs of children with identical or nearly identical test scores. Note: Unless pairings are based on identical scores, there is a possibility that the mean pretest scores of the treatment and comparison groups may differ by amounts large enough to influence the evaluation outcome. If such differences are found, covariance analysis should be used to adjust for them (see Model 2).

Step Four: Once children are paired, randomly assign one member of each pair to the treatment group and the other to the comparison

group. Randomization may be done by flipping a coin, using a table of random numbers, or any other procedure based on chance rather than choice.

Step Five: Once the groups are formed, it is important to monitor their experiences over the treatment period. The experiences of the two groups should be identical with the single exception that one group gets the treatment and the other does not. Where this is not the case, the differences between groups in posttest performance may not be the result of the treatment, but rather a result of uncontrolled attitudinal or experiential factors (see Hazard 10).

Step Six: Administer the posttest. If at all possible, the two groups should be tested at the same time. Large differences in testing times allow potentially relevant experiences to occur for one group and not the other. Even small differences such as the time of day, the weather, the emotional climate and other difficult-to-assess influences may alter test performances.

Step Seven: Score the posttest. Raw scores should be converted to their standard or scale score equivalents before any computations are undertaken. If a test scoring service is used, it should be made clear that each raw score should be converted to its standard or scale score equivalent.

Step Eight: Compute the following summary statistics by obtaining the indicated formulas from any elementary statistics book.

(a) The mean and standard deviation of posttest scores for the treatment group.

(b) The mean and standard deviation of posttest scores for the comparison group.

(c) The correlation between groups based on the original pairing of children.

Note that if one member of a pair is lost, i.e., no posttest score is obtained, the other member must be excluded from all of these calculations. See Chapter VI for further analysis considerations.

Step Nine: Compare the mean posttest scores of the treatment and comparison groups. If the treatment group score is greater than the comparison group score the project may have been effective. The statistical significance of the difference should be checked using the following formula:

$$t_{N-1} = \frac{\bar{Y}_t - \bar{Y}_c}{\sqrt{\frac{s_t^2 + s_c^2 - 2r_{tc}s_t s_c}{N-1}}}$$

where $\bar{Y}_t$ = posttest mean standard score of the treatment group

$\bar{Y}_c$ = posttest mean standard score of the comparison group

s_t = standard deviation of the treatment group posttest scores

s_c = standard deviation of the comparison group posttest scores

r_{tc} = correlation between posttest scores of the two groups

N = number of pairs of children

Degrees of freedom = N-1

The one-tailed probability of the computed t can be found in the tables provided in most standard statistical texts. If it is less than or equal to .05 ($p \leq .05$), the special project may be said to have produced statistically significant achievement gains. There is no generally accepted criterion for deciding whether the size of the gain is large enough to be considered educationally significant. Where standardized tests are used, the standard deviation of the national norm group (σ) provides a useful reference. As a rule of thumb, the authors suggest one-third of a standard deviation above expectation based on posttest scores as a reasonable

cutoff value. In other words, if

$$\bar{Y}_t - \bar{Y}_c \geq \sigma/3$$

the gain may be considered educationally significant.

Model 2

Analysis of Covariance

Summary

General Characteristics. This model is appropriate to use where individual pupils are randomly assigned to treatment and comparison groups or where pre-existing, intact groups which are sufficiently similar to be considered random samples from a single population are assigned to treatment and comparison conditions. Analysis of covariance provides an appropriate statistical adjustment to compensate for pretest score differences between groups if these differences were due to such chance factors as random sampling fluctuations. If pretest differences are real, i.e., the treatment and comparison groups cannot be regarded as random samples from a single population, covariance analysis systematically underadjusts for the initial differences between groups. The underadjustment spuriously reduces the probability that the initially inferior group will be found superior on posttest performance. Conversely, it spuriously increases the probability that the initially superior group will be found superior on posttest performance.

Strengths. When an unmatched control group is used, analysis of covariance provides the best method of adjusting observed posttest scores for random pretest differences. Comparing posttest means that have been adjusted is always more precise than comparing unadjusted posttest scores.

Weaknesses. This model assumes that treatment and comparison students are random samples from a single population so that any difference in pretest performance is due only to sampling error and random error of measurement. It will not provide an appropriate adjustment for pretest score differences which reflect non-random differences between groups (see Hazard 6, page 19). Where analysis of covariance is employed with data from pre-existing, intact groups, there is always some danger in presuming that the groups are random samples from a single population.

Implementation Considerations. This evaluation model allows a wide

choice in the test instruments to be used and in the time of testing. If, as would be recommended, a norm-referenced comparison is also made, the choices become more restricted (see the Norm-referenced Model, page 72).

The control group must be very similar to the treatment group. True random selection is strongly advised. If the groups were not selected randomly, strong evidence is needed to demonstrate that the selection was "random in effect" (see Chapter III, Question 5).

This model involves extensive computations and, unless they can be done at little cost or effort on a computer, a decision should be made as to whether the analysis is justified. The degree of precision gained by employing analysis of covariance depends in part on the correlation between the pretest and the posttest. If the correlation is relatively low, the adjusted values would not differ very much from the unadjusted values; if high, then the posttest means would be adjusted by a correspondingly high proportion of the original pretest difference. Pre- and posttest measures, consequently, should be selected to maximize the correlation between them. Multiple covariates may be used to achieve this objective.

Implementation Procedures.

Step One: Form the treatment and comparison groups. Assignment to groups should be based on a random procedure such as drawing well-shuffled names from a hat. In some cases, intact classrooms may represent a reasonable approximation to randomly selected groups. Groups differing systematically on ethnicity, SES, sex or other obvious variables are never satisfactory. Similarly, a non-volunteer group can never serve as a comparison group for volunteers.

Step Two: Administer and score the pretest. Testing conditions must be exactly alike for the treatment and comparison groups. Testing both groups together may be a good idea unless one group of students is put at a relative disadvantage, e.g., by being tested in unfamiliar surroundings.

Step Three: At the end of the project, administer and score the posttest. Once again, testing conditions for the two groups should

be exactly alike. Raw scores should be converted to their standard or scale score equivalents before any computations are undertaken. If a test scoring service is used, it should be made clear that each raw score should be converted to its standard or scale score equivalent.

Step Four: If there is no difference between the groups on the pretest, analysis of covariance is not needed. In this case, a simple t test for independent groups is appropriate for testing the posttest difference:

$$t_{N_t+N_c-2} = \frac{\bar{Y}_t - \bar{Y}_c}{\sqrt{\left(\frac{N_t s_t^2 + N_c s_c^2}{N_t + N_c - 2}\right)\left(\frac{N_t + N_c}{N_t N_c}\right)}}$$

where $\bar{Y}_t$ = mean standard score of the treatment group on the posttest

$\bar{Y}_c$ = mean standard score of the comparison group on the posttest

s_t = standard deviation of the treatment group posttest scores

s_c = standard deviation of the comparison group posttest scores

N_t = number of treatment group pupils

N_c = number of comparison group pupils

Degrees of freedom = $(N_t + N_c - 2)$

The one-tailed probability of the computed t can be found in the tables provided in most standard statistical texts. If it is less than or equal to .05 ($p \leq .05$), the project may be said to have produced statistically significant achievement gains.

Step Five: Assuming the groups differed in mean pretest scores,

an analysis of covariance is recommended. McNemar (1969, Ch. 18) provides a readable explanation of the model. A more complete development is available in Winer (1971, Ch. 10). Because of the amount of computation involved, the use of a computer is highly desirable. Appropriate programs can be provided by most computer centers. Where the amount of data is small or computer facilities are unavailable, the calculations can be done by hand. Instructions for carrying out the analysis of covariance and a set of worksheets for simplifying the computational work is included in Appendix B. These worksheets are referenced directly to the numerical example in Winer (1971) and preserve his notation, but are revised for the case of two groups (treatment plus comparison). Since the textbook examples are for three groups, they are not directly applicable to the typical project evaluation.

Before undertaking a hand-calculated analysis of covariance it is advisable to do a quick check to see whether the effort is justified. Analysis of covariance is essentially the same as the above *t* test, but with the posttest difference $(\bar{Y}_t - \bar{Y}_c)$ adjusted to take into account differences between the groups at pretest time $(\bar{X}_t - \bar{X}_c)$. If the correlation between pretest and posttest is 1.0 the entire difference $(\bar{X}_t - \bar{X}_c)$ is added to the posttest score of the group which was lower at pretest time. This is the maximum possible adjustment. Since, in practice, the correlation will be less than one, the adjustment will be somewhat smaller. To check whether the adjustment is likely to affect conclusions: (a) test the unadjusted posttest difference $(\bar{Y}_t - \bar{Y}_c)$ using a *t* test, and (b) test the posttest difference with the maximum adjustment using a *t* test. If both *t* tests are significant, then analysis of covariance will also be significant and need not be computed. If both are non-significant, analysis of covariance will also be non-significant and need not be computed. It is only necessary to carry out the analysis of covariance if one *t* test is significant and the other is not. The two *t* tests are as follows:

(a) No adjustment:

Use the above formula for t exactly as written. The no-adjustment numerator is

$$\bar{Y}_t - \bar{Y}_c$$

(b) Maximum adjustment:

Use the above formula for t but change the numerator to

$$(\bar{Y}_t - \bar{X}_t) - (\bar{Y}_c - \bar{X}_c)$$

where $\bar{X}_t$ = mean standard score of the treatment group on the pretest

$\bar{X}_c$ = mean standard score of the comparison group on the pretest

Step Six: Instructions for determining the level of statistical significance for analysis of covariance are included in Appendix B. However, there is no generally accepted criterion for deciding whether the size of the gain is large enough to be considered educationally significant. Where standardized tests are used, the standard deviation of the national norm group (σ) provides a useful reference. As a rule of thumb, the authors suggest one-third of a standard deviation above expectation based on adjusted posttest scores as a reasonable cutoff value. In other words, if

$$\hat{\bar{Y}}_t - \hat{\bar{Y}}_c \geq \sigma/3$$

the gain may be considered educationally significant.

Model 3

Special Regression Models

Summary

General Characteristics. Two special regression models are considered here, the Regression Projection Model (Tallmadge & Horst, 1974) and the Regression-discontinuity Model (Campbell & Stanley, 1963). In both models, the selection of treatment participants is determined on the basis of performance on the pretest. All pupils in a group are pretested and those who score above or below a particular score are assigned to the treatment group while the remaining pupils serve as a comparison group.

Strengths. Both models make use of an identifiable and definable comparison group. This group offers a sounder basis for establishing no-treatment posttest expectations than national norms since the comparability of the experiences of the two groups over the pre-to-posttest interval can be empirically verified. The use of a sharp cutoff score in these models simplifies the interpretation of significant results as compared to regression models which do not require this type of assignment to groups.

Weaknesses. The Regression Projection Model tests the difference between the observed and expected posttest means of the treatment group where the "expectation" is derived from the comparison group regression line. The validity of conclusions based on this model rests on the assumption that the combined-group regression line would be linear over its entire range under no-treatment conditions, an assumption which is not always justified.

The Regression-discontinuity Model tests the difference between the intercepts of the treatment and comparison groups' regression lines with the line representing the pretest cutoff score. In its simplest form this model involves the same assumption of linear regression as does the Regression Projection Model, but by using higher-order regression equations (curved regression lines) the problem can be eliminated (Sween, 1971). A remaining weakness is that where treatment impact is inversely proportional

to pretest scores (i.e., the lowest scoring students make the biggest gains), there may be no difference in regression line intercepts even where the mean gain of the treatment group is highly significant.

Implementation Considerations. Figure 2 on the following page illustrates both the Regression Projection and the Regression-discontinuity Models. In this idealized conception, the solid-line portion of the elipse to the right of the cutoff score represents the actual distribution (scatter plot) of the pre- and posttest scores of the comparison group. It is used to estimate what the score distribution for the treatment group would have been if there had been no special treatment. This no-treatment expectation is illustrated by the broken-line portion of the elipse to the left of the cutoff score. The actual distribution of the treatment group's scores is illustrated by the solid-line portion of the elipse to the left of the cutoff score. This distribution is displaced upward above the no-treatment expected scores indicating that the treatment did have the effect of raising posttest scores.

Regression lines are drawn diagonally through the distributions shown in Figure 2. As mentioned above, the Regression Projection Model involves testing the difference between the observed and the expected mean posttest scores while the Regression-discontinuity Model involves testing the difference between the intercepts of the regression lines with the cutoff score. In the situation shown in Figure 2, regression is linear and the amount of treatment impact was independent of the pupils' pretest scores. Under these conditions, the difference between means is identical to the difference between intercepts and the two evaluation models should yield identical results.

Figure 3 depicts a situation in which the treatment had its greatest impact on pupils farthest below the cutoff score and a negligible effect on pupils right at the pretest cutoff. Under these circumstances, the slope of the treatment group regression line is flatter than that of the comparison group. There is no difference between the intercepts of the regression lines with the cutoff score, but there is a difference between the expected and observed mean posttest scores of the treatment group. While this difference

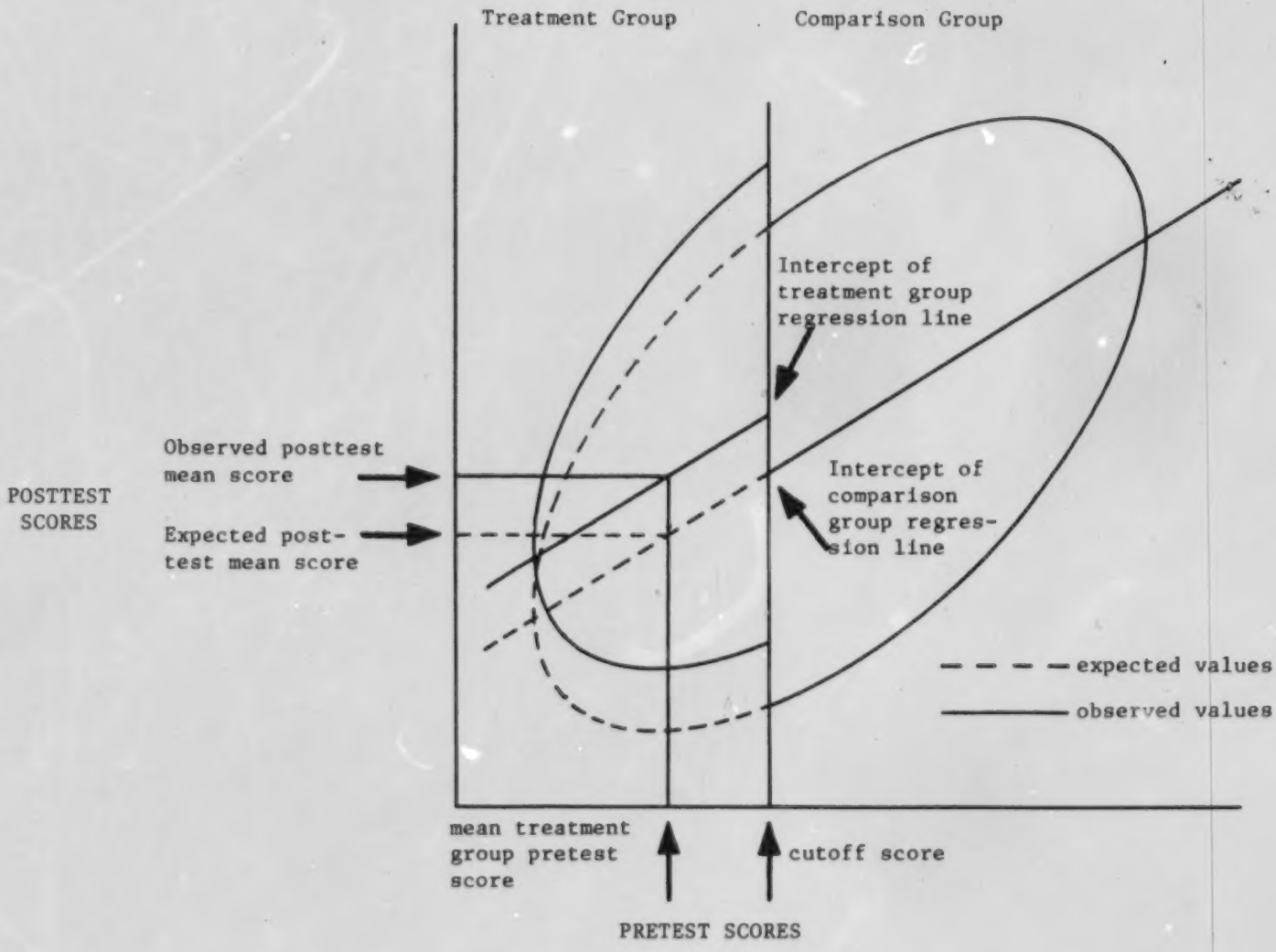

Fig. 2. Score distributions with treatment effect independent of pretest status.

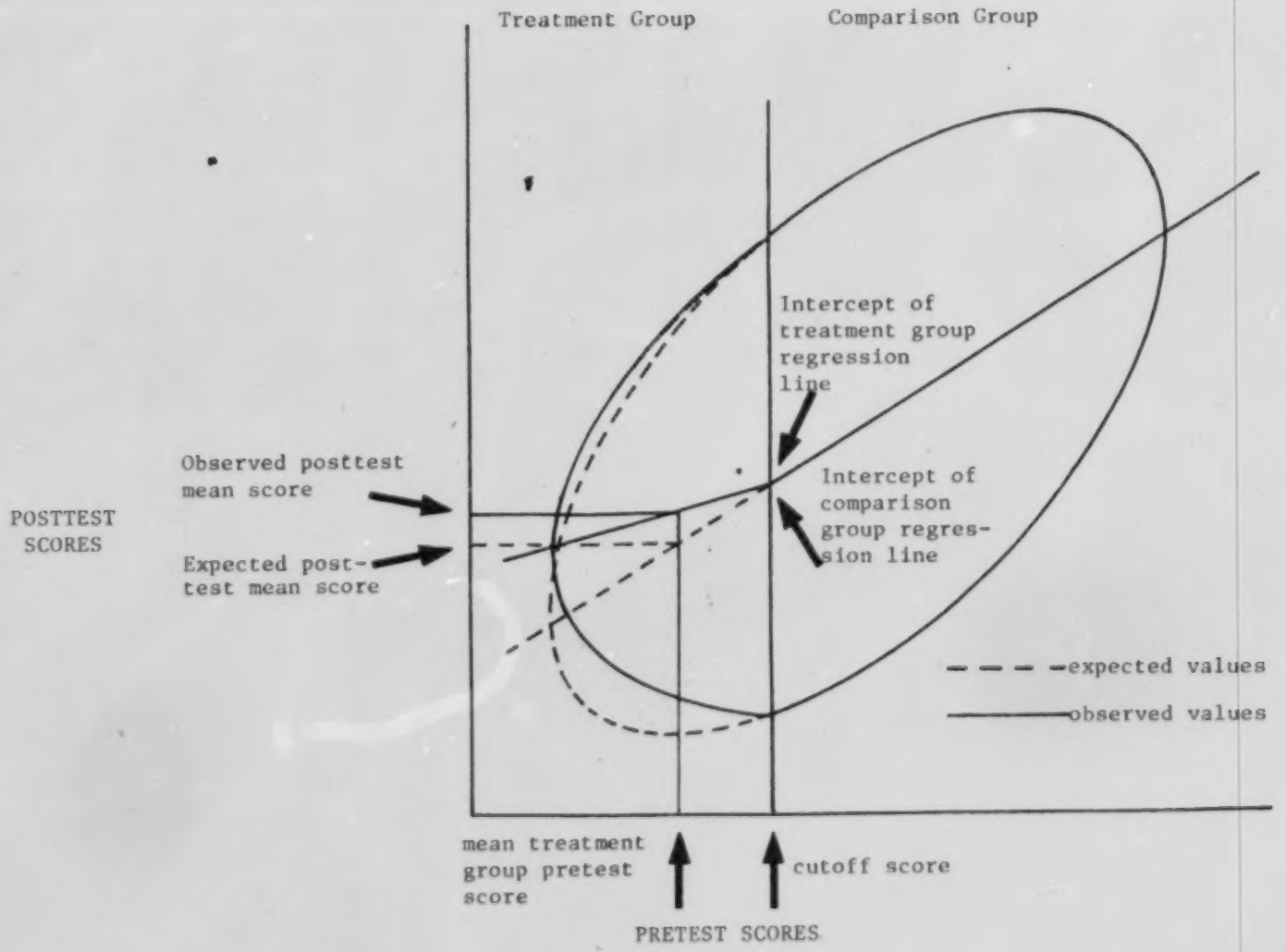

Fig. 3. Score distributions with treatment effect inversely related to pretest status.

would have been detected by the Regression Projection Model and not by the Regression-discontinuity Model, other kinds of treatment impact variations would be more readily detected by the Regression-discontinuity Model. Because its assumptions are less subject to question, the latter model may also be considered more conclusive, especially in its general form (Sween, 1971).

It is not possible to provide general decision criteria as to which of the two models is the more appropriate for a particular situation. Plotting the scatter diagram, however, should provide some insight as to what kinds of influences are operating and, consequently, which of the two models should be used. Knowledge about the treatment may also help. If, for example, a particular project provides remedial instruction in proportion to individual students' needs, it would be more appropriate to expect the kind of impact illustrated in Figure 3 than that shown in Figure 2. In this instance, the Regression Projection Model would be the proper choice.

The utility of both special regression models is proportional to the size of the correlation between pre- and posttest scores. The relationship between the size of this correlation and the validity of inferences which can be drawn from implementation of the models is precisely analogous to the same relationship in Models 1 and 2.

Using test scores as the sole determinant of pupils' needs for special instructional treatments is a practice some educators consider unacceptable. This objection can be resolved by using a composite measure made up, for example, of a pretest score and an independently-made, numerical, teacher rating of need.

One additional point is relevant. The entire discussion of the Regression Projection Model has assumed that the comparison group regression equation would be used to estimate how the treatment group would have performed had they not received the treatment. It would appear equally possible to use the treatment-group regression line to estimate how the comparison group would have performed if they had received the treatment. When the treatment affects the slope of the regression line in the manner shown in Figure 3, however, this practice would lead to the erroneous conclusion that the treatment had a negative impact.

Implementation Procedures.

Step One: To implement either special regression model, administer and score the pretest. The test should be given to all members of a group from which the treatment pupils are to be drawn because of their special needs. The pretest should correlate substantially with the posttest measure.

Step Two: If desired, generate a composite score which incorporates the pretest measure and any other measure such as independently-made teacher ratings.

In generating a composite score, each score element should be weighted in proportion to the standard deviation of scores on that element. For example, a set of pretest scores has a mean of 20 and a standard deviation of 3. These are to be combined with teacher ratings having a mean of 7 and a standard deviation of 5. The pretest scores will thus account for 3/8 of the composite score variability while teacher ratings will account for 5/8 of it. In order to give the two measures equal importance in the composite score, each test score would have to be multiplied by 5/3 to equalize the standard deviations of the two measures. Then the weighted values would simply be summed. Composites involving more than two measures can be constructed in a similar manner although it seems unlikely that a composite involving more than two elements would often be required.

If composite scores are used, it must be remembered that they then become the pretest measure. All future calculations involving "the pretest measure" must use the composite measure--not one of its elements.

Step Three: Establish a single cutoff score. For a remedial project, assign all pupils scoring below this value to the treatment group. Alternately, students scoring above a cutoff score might be assigned to a special project for the gifted. One convenient way to establish a cutoff score is to determine how many pupils can be served by the special project and then count up or

down from the lowest or the highest score until the quota is filled. Once the cutoff is established, it must be adhered to strictly. There can be no exceptions made in the assignment of each grade level if more than one grade level will be involved in the special project.

Step Four: Administer and score the posttest. All available pupils in the original group must be posttested even though only a relatively small proportion of them may have participated in the treatment. The subsequent analyses can be performed using raw scores although it would be preferable to convert both pre- and posttest scores to their standard or scale score equivalents if standardized tests are used.

Step Five: To carry out the computations for either the Regression Projection Model or the simplest version of the Regression-discontinuity Model, calculate the following values:

	Treatment Group	Comparison Group
Number of pupils	N_t	N_c
Mean of (composite) pretest scores	$\bar{X}_t$	$\bar{X}_c$
Standard deviation of (composite) pretest scores	s_{X_t}	s_{X_c}
Mean of posttest scores	$\bar{Y}_t$	$\bar{Y}_c$
Standard deviation of posttest scores	s_{Y_t}	s_{Y_c}
Correlation of posttest with (composite) pretest	r_t	r_c
Slope of the regression line for predicting Y from X	b_t	b_c

Programs are readily available for all computers and programmable calculators to assist in these calculations. The names or descriptions of appropriate programs usually specify that they compute Pearson product-moment correlations and, in general, all of the above values will be printed out automatically.

If no computational facilities are available, the calculations may be done by hand. Computational formulas and instructions may be found in any introductory statistics book. It will simplify the task to recall that

$$b = r \frac{s_Y}{s_X}$$

Once the above values have been calculated the remaining computations are relatively simple.

Step Six A: Regression Projection Model.

In the Regression Projection Model, the actual mean posttest score of the treatment group ($\bar{Y}_t$) is compared with an estimated no-treatment value ($\hat{\bar{Y}}_t$) obtained by projecting the comparison group regression line.

This predicted value is calculated by the following formula:

$$\hat{\bar{Y}}_t = \bar{Y}_c + b_c (\bar{X}_t - \bar{X}_c)$$

The amount of the treatment effect is the difference between the actual and the estimated mean scores, or:

$$\bar{Y}_t - \hat{\bar{Y}}_t$$

The statistical significance of this difference may be tested using the following formula (Tallmadge & Horst, 1974):

$$t_{N_t+N_c-3} = \sqrt{\frac{P_t^2(\bar{Y}_t - \hat{\bar{Y}}_t)^2 \ (N_t + N_c - 3)}{\bar{s}_Y^2 - 2b_c\bar{b}\bar{s}_X^2 + b_c \ \bar{s}_X^2 + P_tP_c(\bar{Y}_t - \hat{\bar{Y}}_t)^2}}$$

Where:

$$P_t = \frac{N_t}{N_t + N_c}$$

$$P_c = \frac{N_c}{N_t + N_c}$$

$$\bar{s}_X^2 = P_t s_{X_t}^2 + P_c s_{X_c}^2$$

$$\bar{s}_Y^2 = P_t s_{Y_t}^2 + P_c s_{Y_c}^2$$

$$\bar{b} = P_t b_t + P_c b_c$$

Step Six: Regression-discontinuity Model

The simplest form of the Regression-discontinuity Model consists of fitting straight regression lines independently to the treatment and comparison groups, then testing the difference between the two lines at the point where they intersect the pretest cutoff score value.

Let:

K = the (composite) pretest cutoff score

$\hat{Y}_t$ = the Y value of the treatment group regression line for a (composite) pretest score of K

$\hat{Y}_c$ = the Y value of the comparison group regression line for a (composite) pretest score of K

Then:

$$\hat{Y}_t = \bar{Y}_t + b_t(K - \bar{X}_t)$$

$$\hat{Y}_c = \bar{Y}_c + b_c(K - \bar{X}_c)$$

Unlike the Regression Projection Model in which a treatment effect was calculated, there is no special interpretation of the value $\hat{Y}_t - \hat{Y}_c$ unless the regression lines have equal slopes. In this case it is a treatment effect. However, if this value is significantly greater than zero, it is evidence of a real treatment effect. The statistical significance of the difference may be tested using the following formula (Sween, 1971):

$$t_{N_t+N_c-4} = \sqrt{\frac{(\hat{Y}_t - \hat{Y}_c)^2 \, (N_t + N_c - 4)}{(V_t + V_c)\left[\frac{N_t + N_c}{N_t N_c} + Z_t + Z_c\right]}}$$

Where:

$$V_t = N_t s_{Y_t}^2 (1 - r_t^2)$$

$$V_c = N_c s_{Y_c}^2 (1 - r_c^2)$$

$$Z_t = \frac{1}{N_t}\left(\frac{K - \bar{X}_t}{s_{x_t}}\right)^2$$

$$Z_c = \frac{1}{N_c}\left(\frac{K - \bar{X}_c}{s_{x_c}}\right)^2$$

The one-tailed probability of the computed t value can be found in the tables provided in most standard statistical texts. The subscripts for t (N_t+N_c-3, or N_t+N_c-4) are the appropriate numbers to use for the "degrees of freedom" column in the table. If P is less than or equal to .05 ($p \leq .05$)

the special project can be said to have produced statistically significant achievement gains.

There is no generally accepted criterion for deciding whether the size of the gain is large enough to be considered educationally significant. Where standardized tests are used, the standard deviation of the national norm group (σ) provides a useful reference for the Regression Projection Model. As a rule of thumb, the authors suggest one-third of a standard deviation as a reasonable cutoff value. In other words, if

$$\bar{Y}_t - \hat{\bar{Y}}_t \geq \sigma/3$$

the gain may be considered educationally significant.

Model 4

General Regression Model

Summary

General Characteristics. This model may be thought of as a more generalized form of the analysis of covariance model. Posttest differences between any two (or more) groups can be tested, adjusting for the effects of any number of quantifiable variables such as pretest scores, sex, SES, location, etc., and their interactions. The effects of using curved regression lines can also be tested and removed.

Strengths. The model itself places no restrictions on the selection of students, their relative pretest performance, or on any facet of the experimental design. Where other models implicitly assume that posttest results are not related to variables other than the pretest, the General Regression Model permits systematic tests of this assumption.

Weaknesses. All forms of the general regression model, including the special case of analysis of covariance, test the hypothesis that posttest differences are the effect of random fluctuations. Where treatment and comparison groups were clearly different to begin with, this is not a useful hypothesis to test (Lord, 1967, p. 38). Regression models are frequently used to statistically "equate" groups which are clearly different but, based on rather plausible assumptions about the nature of the differences, regression models systematically underadjust (Campbell & Erlebacher, 1970). It should be noted that this underadjustment is minor where the correlation between the pretest and the posttest is high and will provide a conservative estimate of project impact where the treatment group has a lower mean pretest score than the comparison group.

Implementation Considerations. While the flexibility of this model may permit an adequate evaluation where none of the other models is feasible, the complexity of both the multivariate statistical manipulations and the experimental design issues create major obstacles to implementation. Only the most sophisticated specialists in these areas

should attempt to plan and implement a study of this nature.

Implementation Procedures. An ad hoc design and detailed procedures must be developed for each evaluation by a qualified specialist. A complete, highly technical, mathematical development of the model is available (Horst, 1974).

Model 5

Norm-referenced Model

Summary

General Characteristics. Project children are compared to a norm group usually comprised of a nationally representative sample of children at the same grade level. The no-treatment expectation is that the project pupils will maintain, at posttesting, the same achievement status with respect to the norm group as they had at pretesting. If their posttest status is higher, the assumption is made that the improvement resulted from participation in the special project.

Strengths. Where no comparison group is available, the norm group provides a plausible estimate of no-treatment posttest scores. Even where a comparison group is available, unless it comes from the same population as the treatment group the Norm-referenced Model offers a more defensible estimate of posttest performance at substantially less cost and effort than a comparison-group design.

Weaknesses. The validity of the model rests on the assumption that the achievement status of a particular subgroup remains constant relative to the norm group over the pre- to posttest interval if no special treatment is provided. Empirical support for this assumption is minimal. It is conceivable that some subgroups would move up and others move down in the normal course of events. When the norm group is like the treatment group, the plausibility of the underlying assumption is greatly enhanced; thus, for example, norms for gifted children would be best for assessing a project serving such pupils.

Implementation Considerations. This model is widely applicable as it does not require a comparison group. The model requires the use of standardized tests. The same level of the same test should be used for both pre- and posttesting (see Hazard 11). Program participants may not be chosen on the basis of their pretest scores (see Hazard 7). Both pre- and posttesting must be accomplished on dates corresponding to the ones on which the test publisher collected normative data (see Hazard 3).

Implementation Procedures.

Step One: Select a standardized achievement test which has real normative data points at dates which are suitable for pre- and posttesting. Information about the normative data points for some of the most commonly used instruments is presented in Appendix A. Similar information about other tests can be derived from the Technical Manuals provided by the test publishers.

It can be seen from Appendix A that most tests have only a single data point either in fall, winter, or spring. Use of these tests requires a 12-month pre- to posttest interval. If high student turnover is expected, it might be better to choose a test for which normative data have been collected in both fall and spring even though the choice of tests is then quite limited.

Step Two: Administer and score the test in exact compliance with the procedures specified by the test publisher. Each test score should be converted to standard or scale scores. If the tests are scored by a scoring service, be sure to specify that each raw score should be converted to its standard or scale score equivalent.

Step Three: Compute the means and standard deviations of the pre- and posttest distributions of the standard or scale scores if these are not provided by a test scoring service. Also compute the correlation between pre- and posttest scores. Computational formulas for these "summary statistics" can be found in any elementary statistics book. It is necessary, of course, to do separate computations for each grade level participating in the project.

Step Four: Look up the percentile equivalents of the mean pretest and posttest standard or scale scores in the norm tables corresponding to the pre- and posttest administration times. The pretest percentile score is used to derive the no-treatment posttest expectation. In the absence of a special treatment, it would

be expected that a group of pupils would maintain its standing relative to the norm group. Thus, the expected posttest score can be found by looking up the standard or scale score equivalent of the pretest percentile in the posttest norms table. This score constitutes the no-treatment posttest expectation.

Step Five: Examine the obtained posttest score in relation to the expected score. If the obtained or observed score is larger than the expected score, there may be some reason to believe that the project was effective. The statistical signficance of the difference should be checked using the following formula:

$$t_{N-1} = \frac{\bar{Y}_{obs} - \bar{Y}_{exp}}{\sqrt{\frac{s_X^2 + s_Y^2 - 2r_{XY}s_Xs_Y}{N - 1}}}$$

Where
- $\bar{Y}_{obs}$ = observed mean posttest score
- $\bar{Y}_{exp}$ = expected mean posttest score
- s_X = pretest standard deviation
- s_Y = posttest standard deviation
- r_{XY} = correlation between pre- and posttest scores
- N = number of children
- $N-1$ = degrees of freedom

The one-tailed probability of the computed t can be found in the tables provided in most standard statistical texts. If p is less than or equal to .05 ($p \leq .05$), the special project may be said to have produced statistically signficant achievement gains.

There is no generally accepted criterion for deciding whether the size of the gain is large enough to be considered educationally significant. Since standardized tests are used, the standard

deviation of the national norm group (σ) provides a useful reference. As a rule of thumb, the authors suggest one-third of a standard deviation above expectation based on posttest scores as a reasonable cutoff value. In other words, if

$$\bar{Y}_{obs} - \bar{Y}_{exp} \geq \sigma/3$$

the gain may be considered educationally significant.

V. GETTING THE DATA (TESTING AND RECORDING)

Once an evaluation design and an appropriate achievement test are chosen, the most crucial step in the evaluation process is the collection of accurate, complete data. Analysis of the data may be a more technically complex step, but when analysis errors are discovered they can usually be corrected. On the other hand, if data are distorted or missing, no amount of analysis can adequately correct the problem. If there are too many flaws in the raw data, the entire evaluation becomes meaningless.

There are four steps in obtaining test data, each requiring planning and decisions: (a) assembling the students, (b) administering the tests, (c) scoring the tests, and (d) recording the scores.

Step 1: Assembling Students for Testing

This step, often passed over lightly, is an important consideration for two reasons. First, of course, the time of day and the place where students are assembled may affect test scores. The date of testing may also be important (see the Norm-referenced Model, page 72). Second, unless the problems are carefully thought out ahead of time, procedures used for pretesting students may prove so cumbersome that changes are made for the posttest. Changes such as testing students in their classrooms rather than in a large assembly hall may or may not make a big difference in scores, but it is certainly not safe to assume that there is no difference. Having to abandon half of a carefully selected control group because posttesting is too expensive is clearly undesirable. Careful planning could avoid all such problems.

It is difficult to generalize about rules for assembling students because of the wide differences among schools. Most important is to minimize the disruption to the students while insuring that all treatment and comparison students can take both pre- and posttests under similar testing conditions. The major problems in achieving this goal are high

absentee rates and distribution of students across a large number of schools. Where the evaluation simply involves testing project students in their regular project setting, few problems should be encountered. If, on the other hand, control students are involved, or if students are to be tested before the project begins or after it ends, then it is well worth the effort to lay out in detail the number of different tests or test levels to be used, the number of test locations, the time for each test, the number of make-up sessions, the number of special test administrators or supervisors, and so on. Testing often turns out to be a bigger project than anticipated, and, if reduction of effort is necessary, it is better to simplify both the pretest and posttest proportionally rather than expending too much effort on the pretest, and then being unable to complete the posttest.

Step 2: Administering the Tests

It goes without saying that test administration should be orderly, and that cheating and other irregularities are not permissible. But orderliness is not enough. For the purposes of evaluation it is necessary to have consistency. There are two kinds of consistency to worry about, depending on whether a norm-referenced or comparison-group evaluation design is used. If a norm-referenced design is used, the critical thing is to be sure that the test publisher's procedures are followed exactly. This specifically includes reading instructions, answering questions, doing practice problems, and timing each section.

When a comparison group is used, it is still advisable to follow the publisher's instructions to the letter so as to make norm-referenced comparison possible, but the most critical thing becomes the similarity between treatment and comparison group testing situations. The most straightforward way of insuring comparable situations is to test both treatment and comparison students as a single group[1] but, usually, in either norm-referenced or comparison-group designs it will be necessary to test several groups, and special steps must be taken to make sure that they are tested under similar conditions so that their scores can be compared.

---

1. However, bringing comparison group pupils into an unfamiliar project lab for testing may put them at a disadvantage.

There are basically two ways of making test situations comparable. One is to use a few carefully trained administrators to test all the groups. The other is to carefully train the regular teachers to give the tests to their own students. At best, the latter alternative is much less desirable from a research viewpoint, and some monitoring of the testing procedures is advisable. If teachers must be used, it may be advisable to have them test each others' classes to minimize possible biases.

Simply telling teachers to look over the test manual is never adequate if one is serious about the evaluation. Each test administrator should be impressed by the importance of following procedures exactly, and each one should have at least "walked through" the entire process, from handing out pencils to collecting the tests, before ever administering the test in an evaluation. Where teacher judgments are involved in scoring student reponses (as in oral reading tests), substantially more training is required.

Step 3: Scoring the Tests

Scoring of standardized tests is usually separate from test administration, so it becomes the third step in the data gathering process. Obviously, the most important requirement in scoring is accuracy, but there are trade-offs of time and money to consider. The major variables are who does the scoring and what type of answer form to use. Most of the major tests can be purchased with machine-scorable booklets or separate answer sheets. Some non-standardized tests may be available only in hand-scored versions.

The main factor in choosing among answer forms is the age of the students. Separate answer sheets are usually much easier to process, but young children tend to score lower on these forms, presumably because the forms are confusing to them. In general, separate answer sheets are suitable for above average fourth graders and all older students. Younger children should use machine-scorable or hand-scored booklets. (Harcourt Brace Jovanovich, Inc., 1973.)

Whichever type of form is used, there are three basic ways of having

the test scored. Scoring can be done by: (a) local school personnel, (b) the publisher of the test, or (c) an independent test scoring company. A choice between the test publisher or an independent company will depend on a variety of variables specific to the local situation and the test that is chosen. Cost, turnaround time, and quality of service may vary, as well as the services offered, and some shopping is in order. The major decision, however, is whether to have the scoring done by either type of service or to simply have the scoring done by available school personnel. Obviously there is no general answer that will apply to all situations. The major advantages of a good scoring service are the accuracy and the variety of analyses provided by computer processing. The major disadvantages are the cost, the care necessary in preparing the answer forms, and the turnaround time. There are also the possibilities that forms will be lost in shipping, or that mishandling or faulty equipment will result in scoring errors. There is little recourse when forms are lost, but spot checks on scoring accuracy should be made after answer forms have been returned.

"Ballpark" figures for machine-scored forms (taken from one widely-used publisher's service) range from $.30 to $.70 per pupil depending on the type of form and length of the test battery. Hand-scored booklets cost three or four times as much to score although a lower original purchase price will offset this difference slightly. Clearly, local personnel can do the basic scoring at lower cost, but included in this publisher's price are a number of features and services that are costly and time consuming when scoring is done by hand. These include: (a) conveniently formated reports in triplicate for each group (e.g., class), completely identified as to test, date, group etc.; (b) raw scores, percentile scores (local or national distributions), and standard scores for each student on each subtest; and (c) mean raw scores for each group. Several other analyses are available for prices ranging from an additional $.05 to $.12 per student for each analysis. These include score distributions for each class, item analyses, and individual student profiles. Additional statistical analyses are readily available, or, for schools with access to their own computer facilities, the scores are available from the publisher on computer cards or tape.

In short, for very small tryouts with simple analyses it may be desirable to do the entire job locally. Unless local computer facilities are available, however, more extensive evaluations may well be completed more accurately, thoroughly, and economically with the help of a scoring service. All the major services have literature and consultants to provide details and to assist in planning the scoring and analysis.

Step 4: Recording the Scores

The final step in data collection is the recording of scores in a usable format. In practice, of course, this may not be a separate step, and certainly planning to record the scores cannot be put off until the other three steps are completed. For example, if a commercial scoring service is used, scores may be returned in the standard format used by the company. This is particularly true when computerized scoring and analyses are done. However, schools that do their own recording or wish to transfer scores from computer printouts to a more convenient form need to think carefully about the way they wish to record their scores. The exact format may seem like a small matter, but in many school districts data from past evaluations are so badly arranged that any analysis (especially where data are filed away for a year or two) is almost impossible. Getting scores copied correctly onto record sheets is not a complicated problem for small-scale local studies, but it must not be overlooked. Even the most conscientious recorders make errors, and all record sheets should be carefully proofread, preferably with one person reading aloud while a second checks the scores.

It is not possible to prescribe a standard format and recording procedure because school requirements and situations vary so much, but some general principles can be established. Basically, scores must be correct, completely identified, and arranged in such a way that they can be easily analyzed. A data form illustrating these principles is shown in Figure 4. Specific characteristics are discussed below, starting at the top of the form.

School_____ Sheet_____of_____

Class/Group_____ Recorder_____

Treatment/Comparison_____ Date_____

Tests:	Pretest	Posttest
Name		
Level		
Form		

Student Names:	ID No.				Date		Raw Score		Stand.Score	
	Cd	Grp	Sx	Ind	Pre	Post	Pre	Post	Pre	Post
1.										
2.										
3.										
4.										
5.										
6.										
7.										
8.										
9.										
10.										
11.										
12.										
13.										
14.										
15.										
16.										
17.										
18.										
19.										
20.										
21.										
22.										
23.										
24.										
25.										

Fig. 4 Sample Data Form

Considerations for data recording forms

1. Most sets of scores require more than one page. The page number identifies each sheet and the "number of pages" helps make sure no pages are missing.

2. Every sheet of paper should have a name and date to indicate who filled in the numbers in case any questions arise in the future.

3. The group that is recorded should be clearly identified at the top of the page to simplify identifying their data when it becomes only one set in a large stack.

4. It simplifies analysis greatly to have only one test (pre and post) recorded on each sheet, provided the rules for listing students (see points 5-10 below) are followed. The complete name of the pretest and posttest (taken exactly from the test booklets and including publication date) must be listed. This point is widely neglected.

5. Identifying students and organizing their names efficiently is the most difficult problem in recording student data. Where evaluations are only for one year and are based on fall and spring testing, the problems can be solved with a little effort and care. But where students must be followed over several years, there is no simple solution since students come and go from projects, and groups are reorganized every year. The simplest rule is to make sure that the posttest scores are all entered on the same sheet of paper as the corresponding pretest scores. This at least eliminates the problem of the evaluator trying to find each student's name on two lists.

6. A second rule for listing student names is to establish a standard ordering of the names, and stick to it for the life of the evaluation and for all tests that are used. If a student moves or fails to take some of the tests, then the appropriate entries should be left blank, but he should not be eliminated from the list. If new students enter the program, their names

should be added to the end of the lists for all tests, even those for which no data will be entered. In addition to the obvious reduction in confusion, there are some practical advantages to this procedure. For example, a master form can be prepared with only the students' names and identification numbers filled in, and the forms can simply be duplicated when new tests are given. It also makes comparisons or correlations between any two sets of scores relatively easy because any two forms can be laid side by side and the corresponding names will line up correctly. If there is a compelling reason to change the order of student names in the middle of a project, then either all forms should be changed or a double set of forms (old and new order) should be maintained.

7. A rule should be established for recording names. "Caldwell, D.E." should never become "Danny Caldwell" on a second list. The simplest procedure is to allow plenty of space and to spell out first names and middle initials (e.g., Caldwell, Daniel E.).
8. Each student should have an ID number that completely identifies him. The example in Figure 4 uses a one-digit experimental condition number, a two-digit group or class identification, a one-digit sex code, and a two-digit student number. In some evaluations, other codes (including letters) can be used, but careful consideration of the situation is necessary in order to permit any desired grouping simply by ID number.
9. The page should be arranged so that it can be photocopied without the students' names. This permits wide use of the data for research purposes without compromising student privacy.
10. A page should have some reasonable number of entries, probably 20 or 25. For some inexplicable reason, numbers like 27 and 33 are popular, and often the number of entries varies from page to page. Unnecessary complications like this help to make the statistician's life miserable.
11. Test dates are critical, especially in norm-referenced evaluations. If all students listed on a page have their pretests in

one day and all are later posttested in a single day, then the test date column is not really necessary. However, this is usually impossible to predict at the time the form is made up, and the columns should be there in order to permit identification of make-up tests and late entries into the program.

12. Pre- and posttest scores should, in general, be in adjacent columns, rather than pairing each pretest raw score with its standard score, percentile score, etc., followed by each posttest score and its transformations. This greatly simplifies the mechanics of analysis; comparisons are nearly always made between pre- and posttest scores of the same type.

VI. ANALYZING THE DATA AND REPORTING THE RESULTS

Analysis

Basic decisions relating to data analysis should be a part of the original evaluation planning. The major decision is the selection of a suitable evaluation model, treated in Chapters III and IV. A second consideration which should be settled at the same time is the division of the students into analysis subgroups. Because of the advantages of having large numbers of students in an analysis, there is some temptation to analyze all available treatment students as a single group, and, where comparison students are used, to combine all of them into a second group. This practice is not justified when distinct subgroups of students are represented. In particular, it is almost never advisable to combine data from (a) different treatment conditions, (b) different grade levels, or (c) different tests. In most education projects it is more meaningful to analyze each subgroup separately, draw separate conclusions for each subgroup, and then summarize the results of these individual analyses. Unless adequate thought is given to the analysis subgroups in the initial planning stages, the subgroups may be too small or too heterogeneous to permit any convincing conclusions.

When the analysis subgroups are determined and the data are in hand, the analysis can proceed. The essential steps for implementing each of the five evaluation models are treated in Chapter IV, but the following preliminary analysis and screening procedures should substantially facilitate interpretation of the formal analysis findings.

A. For students with both pre- and posttest scores:
   (1) Plot the distribution of the pretest raw scores, and compute the mean and standard deviation.
   (2) Plot the distribution of the posttest raw scores, and compute the mean and standard deviation.

(3) Plot the joint pretest-posttest distribution, and compute the product-moment correlation.

B. For students with pretest scores only:

(1) Plot the distribution of the pretest raw scores, and compute the mean and standard deviation.

C. For students with posttest scores only:

(1) These scores are usually not interpretable by themselves, but may be saved for student files or used as baseline data for following-year evaluations.

In general, the size of any achievement gains will be apparent from the above analyses. The differences in mean scores which are tested statistically in the various models can be inspected graphically by comparing the appropriate distributions. However, an equally important use of the plotted distributions is to permit inspection of the data for irregularities which may influence the interpretation of results. It is not possible to list all the kinds of irregularities that might be encountered, but the following occur frequently and are important:

Floor or ceiling effects: Pretest and posttest distributions should be inspected to see whether they are bunched near the top or the bottom of the score range. The top of the score range is simply the highest possible raw score. The bottom of the score range may be zero, but for multiple choice tests it is usually taken to be the score that would be expected if students were simply guessing. For example, in a typical four-choice test students could be expected to get about one fourth of the items correct by guessing. The impacts of floor and ceiling effects are discussed in Hazard 4, page 15.

Large changes in standard deviations from pretest to posttest: A large increase in standard deviation indicates that the project is spreading the students out by helping the initially better students relatively more than the others. A decrease indicates that initially low scoring students are helped relatively more. Either effect would be an important finding and should be described in any evaluation reports on the project.

Low correlations between pre- and posttest scores or irregular joint distributions: These symptoms can be the end result of a variety of problems but, typically, they indicate that the tests are not measuring the attribute of interest with sufficient reliability. If the skill is not measured reliably then, clearly, improvements will not be adequately measured, and positive project results may be obscured. With standardized tests, correlations of .80 to .90 are possible. As correlations drop, results become correspondingly less precise.

Differences between pretested students who took the posttest and those who didn't: If students who have only pretest scores appear to be much different on the pretest from those who took both pre- and posttests, some investigation is required. There are many possible explanations. The better students may graduate, or poorer students may drop out, or both. Such findings are themselves important, and may also be relevant to the interpretation of posttest distributions. If the better students are missing from the posttest distribution, the mean score will be depressed. If the poorer students are missing, the mean score will be spuriously inflated.

Once the data have been carefully examined, the statistical tests of the appropriate model may be applied. In most cases the results will have been clear from inspection of the distributions and a test of significance will serve mainly as a concise, easily reported confirmation that differences were or were not likely to be due to chance factors. It must be remembered that statistical significance depends, in practice, on the number of students in the distributions. Even trivial differences in mean scores become statistically significant when hundreds of students are involved. Conversely, most project effects that could be represented as educationally important will prove to be statistically significant, even with as few as 25 or 30 students.

The question of how big a gain must be before it is considered educationally important is, of course, a judgmental question rather than a statistical one. The evaluator or the project director may well be called upon to offer an opinion on this issue, and while no specific guidelines could cover the variety of settings and situations for educational projects,

the above comments suggest three issues that must be clearly separated in drawing conclusions about the educational importance of project effects. One issue is the size of the project effects. A second is the cost associated with implementing the project, and the third is the conclusiveness of the evaluation results.

The importance of a given project effect usually depends on the cost of the project and the available alternatives. That is, a project that costs very little in money or effort may be very worthwhile even if its effects are rather small, provided there are no obvious, superior alternatives. Any large effect is obviously important in principle, but in practice it may be very costly, and cheaper alternative projects may have comparable effects. While neat cost-effectiveness conclusions are still beyond the state-of-the-art in educational evaluation, decisions should be based on the best information that can be provided.

In addition to the size of the project effect, the conclusiveness of the evaluation should also be discussed. The total evaluation should be weighed in terms of all of the issues discussed in this guidebook, and factors that appear to affect the results should be noted. The hazards discussed in Chapter II, the model weaknesses from Chapters III and IV, and the data collection issues of Chapter V must all be considered. Further, it is the position of the authors that conclusive generalizations about a project are possible only after amassing consistent evidence from a variety of evaluations over a period of time. No single tryout can provide a sound basis for generalizations no matter how carefully it is conducted.

The evaluation of projects within an operational, educational system is an extremely difficult task and decision makers need to become aware of the practical limitations of the process. Often, complications beyond the control of the evaluator preclude any definitive conclusions about project effectiveness, and it is the responsibility of the evaluator to reflect this situation accurately in the evaluation report.

Finally, it should be noted that all of the models in this guidebook are directed at the question of how much better students did in the project

than they would have done without it. Decision makers however, may be interested in some other criterion, such as bringing the mean scores of treatment students up to the national norm. This particular criterion is widely encountered, and while it may represent a meaningful goal, a word of caution is in order. While every evaluator will recognize that exactly half of all students will always be below the national average, it is never safe to assume that the decision maker understands this statistical truism. A brief discussion of the issue, including the reasonableness of the criterion for the particular treatment students in the project, should always accompany any reference to such a criterion.

Reporting

The evaluation report is the final link in the evaluation process. Unless the results are adequately presented, the entire evaluation is of little use to anyone. A variety of people will be interested in the results and, ideally, a separate report should be prepared for each type of audience. In practice, however, only one report will be written and it should cover the requirements of a wide range of readers. The recommendations below assume at least two basic audiences: (a) the local school board and administrators, and (b) educators, government officers, and school personnel outside of the local district. The first group will include non-specialists primarily interested in an easily understood description of the project results. The second group will include skeptical evaluation specialists who must be convinced that the findings are valid. To meet the needs of the first group, a clear summary of the project and the results should be provided. This summary should not be more than two or three pages long and should be included at the front of the report. The body of the report should be concise, but complete, in order to meet the needs of the critical evaluation specialist. It should cover the issues of objectives, costs, and affective changes as well as achievement gains. Report organization and appropriate topics other than achievement gains are discussed in detail in Hawkridge, Campeau, & Trickett (1970). Examples of appropriate section headings and formats can be found in any educational research journal.

In presenting achievement gains, a convincing report must explain

exactly what was done in the evaluation, provide statistics summarizing the results, and justify the conclusions of the evaluators. In preparing the description of what was done, it should be kept in mind that the critical reader will be concerned about all of the hazards in Chapter II of this guidebook and is likely to analyze the evaluation report systematically for possible weaknesses (as in Tallmadge & Horst, 1974). Where information is missing, he will probably assume the worst. Ideally, all of the questions raised in Chapter II and in Tallmadge and Horst (1974), as well as those in Chapter IV specific to a particular model, should be anticipated and discussed.

At a minimum, the report should include a brief description and justification of the model used, a summary of the data, and the results of significance tests. A wide-spread error is the omission of summary statistics that are required if the results are to be meaningful. In particular, evaluation reports often present only mean scores as evidence of effectiveness. While means alone may be sufficient in a report summary, every mean score reported in the body of a report should be accompanied by the number of students represented (N), and the standard deviation of the distribution (s). In addition, it must always be clear whether or not any two means represent exactly the same group of students. Claims of statistical significance should clearly elaborate (or reference) the exact test used, as well as the numerical results of the test. Discussions of educational importance should clearly indicate the local standards against which the project is compared. The local setting also bears on the extent to which the project might be replicable in other school districts, and should be spelled out as clearly as possible.

The evaluator's final decision concerns the saving of information from the evaluation. The published report will provide summarized results, but many of the analyses and statistics recommended in this chapter will not be included. It is not customary, for example, to include graphs of score distributions in a report unless they illustrate some special point. Most evaluators will, however, want to keep these graphs plus all calculated statistics on file for future reference. Whether the raw data recording sheets are saved or not depends on local policy and on the possible use of

the data in evaluations during subsequent years. Providing the preliminary analyses of this chapter and the specific analyses of Chapter IV have been carefully completed and documented, it is unlikely that the raw data will be needed for future reanalyses.

APPENDIX A

Characteristics of Commonly Used Standardized Tests

Page

1. California Achievement Test (1970 Edition). 93
2. Cooperative Primary Tests (1965 Edition) 95
3. Comprehensive Test of Basic Skills (1968 Edition) 96
4. Gates-MacGinitie Reading Tests (1964 Edition) 98
5. Iowa Test of Basic Skills (1971 Edition) 100
6. Metropolitan Achievement Tests (1970 Edition) 102
7. Sequential Tests of Educational Progress II (1969 Edition). . . . 103
8. SRA Achievement Series (1971 Edition) 105
9. Stanford Achievement Tests (1973 Edition) 107

1. California Achievement Test (1970 Edition)

A. Levels/Grades/Forms

Level 1 / Grades 1.5-2 / Form A
Level 2 / Grades 2-4 / Form A
Level 3 / Grades 4-6 / Form A
Level 4 / Grades 6-9 / Form A
Leve 5 / Grades 9-12 / Form A

B. Normative Data Point

February-March (beginning- and end-of-year norms are projections and should not be used in norm-referenced evaluations.)

C. Types of Scores

Raw Scores (appropriate for use with Anchor Test Study Equivalency Tables)

Grade-equivalent Scores

Achievement Development Scale Scores (expanded standard scores) (should be used for all statistical computations not involving Anchor Test Study conversions)

Percentiles and Stanines (beginning- and end-of-year scores are projections and should not be used in norm-referenced evaluations)

D. Comments

The reading scales of Levels 3 (Grades 4 and 5) and 4 (Grade 6) were included in the Anchor Test Study. The CAT may thus be used for norm-referenced evaluations under the following conditions:

1. Pretest and posttest in late February (12-month interval) using CAT norms

2. Pretest and posttest in mid-April (12-month interval) using Anchor Test Study Individual Score Norm*. Reading only, and grades 4, 5, and 6 only.

3. Pretest in mid-October, posttest in mid-April Using Anchor Test Study Equivalency Tables* and Metropolitan Achievement Test norms. Reading only, and grades 4, 5, and 6 only.

---

* The following procedure is recommended for use with Anchor Test Study data. First, convert each pupil's CAT raw score to the equivalent MAT raw score. Second, convert each MAT raw score to its corresponding standard score. Third, calculate all statistics using MAT standard scores. Then, if Anchor Test Study norms are to be used, convert the mean MAT standard score to its MAT raw score equivalent. The corresponding percentile can then be read out of the Individual Score Norms Tables (not the School Means Norms Tables). If the MAT norms are to be used, percentile equivalents are provided corresponding to mean standard scores.

2. Cooperative Primary Tests (1965 Edition)

A. Levels/Grades/Forms

12 / Grades 1.5-2.0 / Forms A & B

23 / Grades 2.0-3.9 / Forms A & B

B. Normative Data Points

Late October-early November and late April-early May

C. Types of Scores

Raw Scores

Scale Scores (expanded standard scores) (should be used for all statistical computations)

Percentiles

D. Comments

This test has appropriate norms for a fall pretest-spring posttest norm-referenced evaluation. It was not included in the Anchor Test Study because it does not cover grades 4, 5, and 6.

3. Comprehensive Test of Basic Skills (1968 Edition)

A. Levels/Grades/Forms

Level 1 / Grades 2.5-4 / Forms Q & R
Level 2 / Grades 4-6 / Forms Q & R
Level 3 / Grades 6-8 / Forms Q & R
Level 4 / Grades 8-10 / Forms Q & R

B. Normative Data Point

Last week of February-first week of March (Beginning- and End-of-year norms are projections and should not be used in norm-referenced evaluations.

C. Types of Scores

Raw Scores (appropriate for use with Anchor Test Study Equivalency Tables)
Grade-equivalent Scores
Expanded Standard Scores (should be used for all statistical computations not involving Anchor Test Study conversions)
Percentiles and Stanines (Beginning- and End-of-year scores are projections and should not be used in norm-referenced evaluations.

D. Comments

The reading scales of Level 2, Form Q (Grades 4 and 5) and Level 3, Form Q (Grade 6) were included in the Anchor Test Study. The CTBS may thus be used for norm-referenced evaluations under the following conditions:

1. Pretest and posttest at end of February-beginning

of March (12-month interval) using Anchor Test Study Individual Score Norms* in reading only, and grades 4, 5, and 6 only.

3. Pretest in mid-October, posttest in mid-April using Anchor Test Study Equivalency Tables* and Metropolitan Achievement Test norms. Reading only, and grades 4, 5, and 6 only.

---

* Procedures recommended for using Anchor Test Study Equivalency Tables and norms with the California Achievement Test are presented in the footnote on page 94. The same procedures should be used with Form Q of the CTBS. If Form R of the CTBS is used, each raw score must be converted to its Form Q equivalent (using conversion tables provided by the publisher) before the Anchor Test Study tables are used.

4. Gates-MacGinitie Reading Tests (1964 Edition)

A. Levels/Grades/Forms

Primary A / 1.5-2.0 / 1, 1M, 2, 2M
Primary B / 2.0-3.0 / 1, 1M, 2, 2M
Primary C / 3.0-4.0 / 1, 1M, 2, 2M
Primary CS / 2.5-4.0 / 1, 1M, 2, 2M, 3, 3M
Survey D / 4.0-7.0 / 1M, 2M, 3M
Survey E / 7.0-10.0/ 1M, 2M, 3M

B. Normative Data Points

October and April except January for first grade. (February and May norms are projections. Because of the proximity of the May norms to the April data point, the May norms are probably adequate for use with norm-referenced comparisons. The February norm, however, cannot be recommended for use with such comparisons.)

C. Types of Scores

Raw Scores (appropriate for use with Anchor Test Study Equivalency Tables)
Grade Scores
Standard Scores (should be used for all statistical computations not involving Anchor Test Study conversions)

D. Comments

The standard scores provided for the Gates-MacGinitie are not expanded standard scores. It is thus not possible to relate scores from one level of the test to norms for another level, and using test levels with appropriate norms may produce ceiling or floor effects when disadvantaged or gifted students are tested. (See Hazard 4, p. 15.)

Survey D, Form 1M was included in the Anchor Test Study. The Gates-MacGinitie may thus be used for norm-referenced evaluations under the following conditions:

1. Pretest in mid-October, posttest in mid-May using Gates-MacGinitie norms (but with the possibility that ceiling and floor effects may be encountered)
2. Pretest and posttest in mid-April (12-month interval) using Anchor Test Study Individual Score Norms*. Grades 4, 5, and 6 only.
3. Pretest in mid-October and posttest in mid-April using Anchor Test Study Equivalency Tables* and Metropolitan Achievement Test norms. Grades 4, 5, and 6 only.

---

* Procedures recommended for using Anchor Test Study Equivalency Tables and norms with the California Achievement Test are presented in the footnote on page 94. The same procedures should be used with Form 1M of the Gates-MacGinitie. The implication of using other forms is not clear as score equivalency tables are not provided by the publishers despite the probable existence of between-form differences. The test publishers apparently presume that the differences are so small as to be negligible.

5. Iowa Test of Basic Skills (1971 Edition)

A. Levels/Grades/Forms

Level 7 / 1.7-2.5 / Forms 5 & 6
Level 8 / 2.6-3.5 / Forms 5 & 6
Level 9 / 3.0-3.9 / Forms 5 & 6
Level 10 / 4.0-4.9 / Forms 5 & 6
Level 11 / 5.0-5.9 / Forms 5 & 6
Level 12 / 6.0-6.9 / Forms 5 & 6
Level 13 / 7.0-7.9 / Forms 5 & 6
Level 14 / 8.0-8.9 / Forms 5 & 6

B. Normative Data Point

Last half of October, first half of November (Mid-year and Spring norms are projections and should not be used for norm-referenced evaluations)

C. Types of Scores

Raw Scores (appropriate for use with Anchor Test Study Equivalency Tables)
Grade-equivalent Scores
Age-equivalent Scores
Standard Scores (expanded standard scores) (should be used for all statistical computations not involving Anchor Test Study conversions)
Percentiles and Stanines (Mid-year and Spring scores are projections and should not be used for norm-referenced evaluations)

D. Comments

The reading scales of Levels 10 (Grade 4), 11 (Grade 5), and 12 (Grade 6), Form 5 were included in the Anchor Test Study. The ITBS may thus be used for norm-referenced evaluation under the following conditions:

1. Pretest and posttest in late October-early November (12-month interval) using ITBS norms

2. Pretest and posttest in mid-April (12-month interval) using Anchor Test Study Individual Score Norms.* Reading only, and grades 4, 5, & 6 only.

3. Pretest in mid-October and posttest in mid-April using Anchor Test Study Equivalency Tables* and Metropolitan Achievement Test norms. Reading only, and grades 4, 5, & 6 only.

---

* Procedures recommended for using Anchor Test Study Equivalency Tables and norms with the California Achievement Test are presented on page 94. The same procedures should be used with Form 5 of the ITBS. The implications of using other forms is not clear as score equivalency tables are not provided despite the fact that some between-form differences are present. The test publishers apparently presume that the differences are so small as to be negligible.

6. Metropolitan Achievement Tests (1970 Edition)

A. Levels/Grades/Forms

Primary 1	/ 1.5-2.4 /	F, G, H
Primary 2	/ 2.5-3.4 /	F, G, H
Elementary	/ 3.5-4.9 /	F, G, H
Intermediate	/ 5.0-6.9 /	F, G, H
Advanced	/ 7.0-9.5 /	F, G, H

B. Normative Data Points

Mid-October and mid-April

C. Types of Scores

Raw Scores
Grade-equivalent Scores
Standard Scores (expanded standard scores) (should be used for all statistical computations)
Percentiles and Stanines

D. Comments

The reading scales of Form F of the Elementary (Grade 4) and Intermediate (Grades 5 and 6) Levels were included in the Anchor Test Study. The MAT may thus be used for norm-referenced evaluation under the following conditions:

1. Pretest in mid-October and posttest in mid-April using MAT norms.
2. Pretest and posttest in mid-April (12-month interval) using Anchor Test Study Individual Score Norms*. Reading only, and grades 4, 5, and 6 only.

* If Anchor Test Study norms are to be used, convert the mean MAT standard score to its raw score equivalent. The corresponding percentile can then be read out of the Individual Score Norms Table (not the School Means Norms Tables). If the MAT norms are to be used, percentile equivalents are provided corresponding to mean standard scores.

7. Sequential Tests of Educational Progress II (1969 Edition)

A. Levels/Grades/Forms

4 / 4-6 / A, B
3 / 7-9 / A, B
2 / 10-12 / A, B

B. Normative Data Point

Last week in April, first three weeks in May (Fall norms are identical to the spring norms for the previous grade. As such, they should not be used in norm-referenced evaluations.

C. Types of Scores

Raw Scores (appropriate for use with Anchor Test Study Equivalency Tables)
Converted Scores (expanded standard scores) (should be used for all statistical computations not involving Anchor Test Study conversions)
Percentiles and Stanines (Fall scores are projections and should not be used in norm-referenced evaluations)

D. Comments

The reading scales of Level 4, Form A, were included in the Anchor Test Study. STEP II may thus be used for norm-referenced evaluations under the following conditions:

1. Pretest and posttest in early May (12-month interval) using STEP II norms.
2. Pretest and posttest in mid-April (12-month interval) using Anchor Test Study Individual Score Norms*. Reading only, and grades 4, 5, & 6 only.

---

* Procedures recommended for using Anchor Test Study Equivalency Tables and norms with the California Achievement Test are presented in the footnote

3. Pretest in mid-October, posttest in mid-April using Anchor Test Study Equivalency Tables* and Metropolitan Achievement Test norms. Reading only, and grades 4, 5, & 6 only.

---

on page 94. The same procedures should be used with Form A of STEP II. If Form B is used, each raw score must be converted to its Form A equivalent (using conversion tables provided by the publisher) before the Anchor Test Study Tables are used.

8. SRA Achievement Series (1971 Edition)

A. Levels/Grades/ Forms

Primary I	/ 1.0-5.5 /	E, F
Primary II	/ 1.0-5.9 /	E, F
Blue	/ 3.5-8.5 /	E, F
Green	/ 4.5-9.9 /	E, F
Red	/ 6.5-10.5/	E, F

B. Normative Data Point

Mid-April (Beginning- and middle-of-year norms are projections and should not be used in norm-referenced evaluations)

C. Types of Scores

Raw Scores (appropriate for use with Anchor Test Study Equivalency Tables)
Grade-equivalent Scores
Growth Score Values
Percentiles and Stanines (Beginning- and Middle-of-year scores are projections and should not be used in norm-referenced evaluations)

D. Comments

Form E of the Blue level (Grades 4 and 5) and the Green level (Grade 6) were included in the Anchor Test Study. The SRA Achievement Tests may thus be used for norm-referenced evaluations under the following conditions:

1. Pretest and posttest in mid-April (12-month interval) using SRA Achievement norms.
2. Pretest and posttest in mid-April (12-month interval) using Anchor Test Study Individual Score Norms*. Reading only, grades 4, 5, & 6 only.

---

* Procedures recommended for using Anchor Test Study Equivalency Tables and norms with the California Achievement Test are presented in the footnote

3. Pretest in mid-October and posttest in mid-April using Anchor Test Study Equivalency Tables* and Metropolitan Achievement Test norms. Reading only, grades 4, 5, & 6 only.

---

on page 94. The same procedures should be used with Form E of the SRA Achievement Tests. The implication of using Form F is not clear as score equivalency tables are not provided by the publishers despite the probable existence of between-form differences. The test publishers apparently presume that the differences are so small as to be negligible.

9. Stanford Achievement Tests (1973 Edition)

A. Levels/Grades/Forms

Primary I	/ 1.5-2.8 /	A, B, C
Primary II	/ 2.5-3.8 /	A, B, C
Primary III	/ 3.8-4.8 /	A, B, C
Intermediate I	/ 4.8-5.8 /	A, B, C
Intermediate II	/ 5.8-7.8 /	A, B, C
Advanced	/ 7.1-9.8 /	A, B, C

B. Normative Data Points

October, February, and May (Most of the SAT percentile and stanine norms tables are closely tied to empirical data. The following, however, are projections and should not be used for norm-referenced evaluations: Primary II, grade 3.5; Primary III, grades 3.5 and 4.5; Intermediate I, grades 4.5 and 5.5; Intermediate II, grades 6.5 and 7.5; Advanced, grades 7.5, 8.5, and 9.5.)

C. Types of Scores

Raw Scores

Grade-equivalent Scores

Scaled Scores (expanded standard scores) (should be used for all statistical computations)

Percentiles and Stanines (percentiles and stanines obtained from the projected norms tables listed above should not be used for norm-referenced evaluations.)

D. Comments

An earlier edition of the Stanford Achievement Tests (1964) was included in the Anchor Test Study. The new edition, however, has many advantages over the old and should be preferred despite the fact that it cannot be used in conjunction with the Anchor Test Study Equivalency Tables.

APPENDIX B

Analysis of Covariance Worksheets

Analysis of covariance is both theoretically and computationally complex. An evaluator undertaking this analysis should have access to a good reference book describing the approach in detail. Tatsuoka (1971, Ch. 3) and McNemar (1969, Ch. 18) provide readable explanations of the model. A more complete development is available in Winer (1971, Ch. 10). Because of the amount of computation involved, the use of a computer is highly desirable. Appropriate programs can be provided by most computer centers.

Where the amount of data is small or computer facilities are unavailable, the calculations can be done by hand. This appendix provides a set of worksheets for simplifying the computational work. The worksheets are referenced directly to the numerical example in Winer (1971, p. 775) and preserve his notation, but are revised for the case of two groups (treatment plus comparison). Since the textbook example is for three groups, it is not directly applicable to the typical project evaluation.

Four worksheets are provided:

Worksheet One is used to record intermediate results that are used for the remaining calculations. All of the terms in columns one and two will be available from the preliminary analyses recommended in Chapter VI.

Worksheet Two is used to arrive at the basic test of significance of the project effects.

Worksheet Three is used to test whether the regression lines for the two groups have the same slope. If the F ratio for the regression lines is significant (i.e., the two slopes are not equal) then analysis of covariance should not be used, and the F ratio from Worksheet Two is meaningless. Logically, Worksheet Three should be completed before

Worksheet Two. Only items (202), (205), and (208) from Worksheet Two are needed to complete Worksheet Three.

Worksheet Four is used to calculate the adjusted mean posttest scores. These adjusted scores are used only to provide an estimate of the "real" effect of the project. They may be useful in determining "educational" significance, but are not involved in the computation of statistical significance.

Significance Levels:

Tables of F values are available in McNemar (1969, pp 509-511) and Winer (1971, pp 864-868). In McNemar:

n_1 = degrees of freedom (df) for the numerator

n_2 = degrees of freedom (df) for the denominator

The .05 level of significance is suggested in this guidebook. Winer uses the notation $(1-\alpha) = .95$ for the .05 level of significance.

Notation Used on the Worksheets:

i = student number

j = group ID (i.e., j = Treatment (t) or Comparison (c))

X_{ij} = pretest raw score for student i of group j

Y_{ij} = posttest raw score for student i of group j

n_j = number of students in group j

N = total number of students ($N = n_t + n_c$)

The remaining notation in this appendix follows Winer (1971). It may be helpful to note that in Winer:

S refers to "total" variation

E refers to "error" or "within group" variation

T refers to "treatment" or "between groups" variation

and that on page 775 (omitting subscripts):

$$(1x) = \frac{(\Sigma\Sigma X)^2}{N} \qquad (1xy) = \frac{(\Sigma\Sigma X)(\Sigma\Sigma Y)}{N} \qquad (1y) = \frac{(\Sigma\Sigma Y)^2}{N}$$

$$(2x) = \Sigma\Sigma X^2 \qquad (2xy) = \Sigma\Sigma XY \qquad (2y) = \Sigma\Sigma Y^2$$

$$(3x) = \sum_j \left(\frac{(\Sigma X)^2}{n_j} \right) \qquad (3xy) = \sum_j \left(\frac{(\Sigma X)(\Sigma Y)}{n_j} \right) \qquad (3y) = \sum_j \left(\frac{(\Sigma Y)^2}{n_j} \right)$$

The double summation signs ($\Sigma\Sigma$) indicate that the values are first summed over all n_j students in each group, then the two group sums are added together.

On all worksheets, results which are needed for later calculations are identified by a three-digit number. The number (148), for example, indicates Worksheet One, Column 4, Row 8. Worksheets Two and Three are not divided into columns, so, for example, (212) indicates Worksheet Two, Row 12.

There are no mathematical checks built into the worksheets. To insure accuracy it is <u>essential</u> to have two persons complete the calculations independently.

ANALYSIS OF COVARIANCE

Worksheet One (Winer, 1971, pp. 775-776)

Sums, Sums of Squares, Sums of Crossproducts

	j	Column 1 Treatment Group	Column 2 Comparison Group	Column 3 (Col 1 + Col 2)	Column 4 Intermediate Results	
Pretest Scores	n_j	_____(110)	_____(120)	_____(130)	N = (130)	: _____(140)
	ΣX	_____(111)	_____(121)	_____(131)	(1x) = $(131)^2/(130)$	: _____(141)
	ΣX^2	_____(112)	_____(122)	_____(132)	(2x) = (132)	: _____(142)
	$(\Sigma X)^2/n_j$	_____(113)	_____(123)	_____(133)	(3x) = (133)	: _____(143)
Posttest Scores	ΣY	_____(114)	_____(124)	_____(134)	(1y) = $(134)^2/(130)$	: _____(144)
	ΣY^2	_____(115)	_____(125)	_____(135)	(2y) = (135)	: _____(145)
	$(\Sigma Y)^2/n_j$	_____(116)	_____(126)	_____(136)	(3y) = (136)	: _____(146)
	$(\Sigma\Sigma X)(\Sigma\Sigma Y)/N$	——————————————————————————>			(1xy)= (131)(134)/(130)	: _____(147)
	ΣXY	_____(118)	_____(128)	_____(138)	(2xy)= (138)	: _____(148)
	$(\Sigma X)(\Sigma Y)/n_j$	_____(119) [(111)(114)/(110)]	_____(129) [(121)(124)/(120)]	_____(139)	(3xy)= (139)	: _____(149)

ANALYSIS OF COVARIANCE

Worksheet Two (Winer, 1971, pp. 775-778)

Computation of F ratio for the significance of the

adjusted difference between the Treatment and Comparison groups

S_{xx} = (142) - (141) : _____ - _____ = _____ (201)

E_{xx} = (142) - (143) : _____ - _____ = _____ (202)

T_{xx} = (201) - (202) : = _____ (203)

S_{xy} = (148) - (147) : _____ - _____ = _____ (204)

E_{xy} = (148) - (149) : _____ - _____ = _____ (205)

T_{xy} = (204) - (205) : _____ (206)

S_{yy} = (145) - (144) : _____ - _____ = _____ (207)

E_{yy} = (145) - (146) : _____ - _____ = _____ (208)

T_{yy} = (207) - (208) : _____ (209)

S'_{yy} = (207) - $(204)^2/(201)$: _____ - _____ = _____ (210)

E'_{yy} = (208) - $(205)^2/(202)$: _____ - _____ = _____ (211)

T_{yyR} = (210) - (211) : _____ (212)

$$F = \frac{T_{yyR}/(df_T)}{E'_{yy}/(df_{E'})}$$

$$= \frac{(212)[(130)-3]}{(211)} : \frac{[_____][_____]}{[\quad]} = _____ \quad (213)$$

degrees of freedom: $\frac{1}{\Sigma(n_j-1)-1} = \frac{1}{N-3}$ numerator / denominator

ANALYSIS OF COVARIANCE

Worksheet Three (Winer, 1971, pp. 777-778) Test for Homogeneity of Within-group Regression

$E_{xy_j} = \Sigma X_j Y_j - (\Sigma X_j \Sigma Y_j)/n_j$

$E_{xy_t} = (118) - [(111)(114)]/(110)$: _____ - [_____][_____]/_____ = _____ (301)

$E_{xy_c} = (128) - [(121)(124)]/(120)$: _____ - [_____][_____]/_____ = _____ (302)

$E_{xx_j} = \Sigma X_j^2 - (\Sigma X_j)^2/n_j$

$E_{xx_t} = (112) - (113)$: _____ - _____ = _____ (303)

$E_{xx_c} = (122) - (123)$: _____ - _____ = _____ (304)

$\sum_j E_{xy_j}^2/E_{xx_j} = \frac{(301)^2}{(303)} + \frac{(302)^2}{(304)}$: $\frac{[\quad]^2}{[\quad]} + \frac{[\quad]^2}{[\quad]}$ = _____ (305)

$S_1 = E_{yy} - \sum_j E_{xy_j}^2 / E_{xx_j} = (208) - (305)$: _____ - _____ = _____ (306)

$S_2 = \sum \frac{E_{xy_j}^2}{E_{xx_j}} - \frac{E_{xy}^2}{E_{xx}} = (305) - \frac{(205)^2}{(202)}$: _____ - $\frac{[\quad]^2}{[\quad]}$ = _____ (307)

$F = \frac{S_2 / (1)}{S_1/\Sigma(n_j-2)} = \frac{(307)}{(306)}\left[(130)-4\right]$: $\frac{[\quad][\quad]}{[\quad]}$ = _____ (308)

Degrees of freedom: 1 numerator
N-4 denominator

ANALYSIS OF COVARIANCE

Worksheet Four (Winer, 1971, p. 779)

Computation of "Adjusted" Mean Posttest Scores

Slope of the Within-group Regression Line:

$$b = \frac{E_{xy}}{E_{xx}} = \frac{(205)}{(202)} : \frac{[\qquad\qquad]}{[\qquad\qquad]} = \qquad\qquad (401)$$

	Treatment Group	Comparison Group	Total Group
$\bar{X}_j$	(111)/(110) : _____(412)	(121)/(120) : _____(422)	(131)/(130) : _____(432) = $\bar{X}$
$\bar{X}_j - \bar{X}$	(412)-(432) : _____(413)	(422)-(432) : _____(423)	
$\bar{Y}_j$	(114)/(110) : _____(414)	(124)/(120) : _____(424)	
$b(\bar{X}_j - \bar{X})$	(401)(413) : _____(415)	(401)(423) : _____(425)	
$\bar{Y}' = \bar{Y}_j - b(\bar{X}_j - \bar{X})$	(414)-(415) : _____(416)	(424)-(425) : _____(426)	

REFERENCES

Campbell, D. T. Making the case for randomized assignment to treatments by considering the alternatives: Six ways in which quasi-experimental evaluations in compensatory education tend to underestimate effects. Draft report # Pre627, Northwestern University, Evanston, Illinois, June, 1974.

Campbell, D. T., & Erlebacher, A. E. How regression artifacts in quasi-experimental evaluations can mistakenly make compensatory education look harmful. In J. Hellmuth (Ed.), Disadvantaged child. Vol. 3. Compensatory education: A national debate. New York: Bruner/Mazel, 1970.

Campbell, D. T., & Stanley, J. C. Experimental and quasi-experimental designs for reserach on teaching. In N. L. Gage (Ed.), Handbook of research on teaching. Chicago: Rand McNally, 1963. (Also published as Experimental and quasi-experimental designs for reserach. Chicago: Rand McNally, 1966.)

Cronbach, L. J. Essentials of psychological testing (Third edition). New York: Holt, Rinehart and Winston, Inc., 1968.

Della-Piana, G. M. Reading diagnosis and prescription: An introduction. New York: Holt, Rinehart and Winston, Inc., 1968.

Foat, C. M. Selecting exemplary compensatory education projects for dissemination via project information packages. Los Altos, Calif.: RMC Research Corporation, May, 1974. (Technical Report No. UR-242)

Harcout Brace Jovanovich, Inc. The effect of separate answer document use on achievement test performance of grade 3 and 4 pupils. New York: Metropolitan Achievement Test Special Report No. 24., June, 1973.

Hawkridge, D. G., Campeau, P. L., & Trickett, P. K. Preparing evaluation reports: A guide for authors. Pittsburgh: American Institutes for Research. 1970. (AIR Monograph No. 6.) (Also published under the same title as OE-10065 by U. S. Government Printing Office, Washington, D. C.: 1970.)

Horst, P. Effect of treatment as a special case of generalized multiple regression. Eugene, Oregon: Oregon Research Institute, 1974. (ORI Technical Report Vo. 14, No. 2.)

Lord, F. M. Elementary models for measuring change. In C. W. Harris (Ed.), Problems in measuring change. Madison, Wisconsin: University of Wisconsin Press, 1967.

Loret, P. G., Seder, A., Bianchini, J. C., & Vale, C. A. Anchor test study: Equivalence and norms tables for selected reading achievement tests (grades 4, 5, 6). Office of Education Report 74-305, U. S. Government Printing Office, Washington, D. C.: 1974.

McNemar, Q. Psychological statistics (Fourth edition). New York: John Wiley and Sons, Inc., 1969.

Saretsky, G. The OEO P. C. experiment and the John Henry effect. Phi Delta Kappan, 1972, 53, 579-581.

Sween, J. A. The experimental regression design: An inquiry into the feasibility of nonrandom treatment allocation. Unpublished doctoral dissertation, Northwestern University, 1971.

Tallmadge, G. K. The development of project information packages for effective approaches in compensatory education. Los Altos, Calif.: RMC Research Corporation, October, 1974 (Technical Report No. UR-254)

Tallmadge, G. K. & Horst, D. P. A procedural guide for validating achievement gains in educational projects. Los Altos, Calif.: RMC Research Corporation, May 1974 (Technical Report No. UR-240)

Tatsuoka, M. M. Multivariate analysis: Techniques for educational and psychological research. New York: John Wiley and Sons, Inc., 1971.

Tyler, L. E. Human abilities. in P. H. Mussen & M. R. Rosenzweig (Eds.) Annual review of psychology. Palo Alto, Calif.: Annual Reviews Inc., 1972.

Whitehead, T. N. The industrial worker. Vol. 1. Cambridge: Harvard University Press, 1938.

Winer, B. J. Statistical principles in experimental design. (Second edition) New York: McGraw-Hill, 1971.